Xây dựng Hệ thống Dịch máy Anh-Việt: Từ Kiến trúc Transformer Nguyên bản đến Tối ưu hóa LLM trong Lĩnh vực Y tế

Vũ Văn Phong
Khoa Công nghệ Thông tin
Đại học Công Nghệ - Đại Học Quốc Gia Hà Nội
22028309

Triệu Việt Hùng
Khoa Công nghệ Thông tin
Đại học Công Nghệ - Đại Học Quốc Gia Hà Nội
22028069

Đinh Huyền Trang
Khoa Công nghệ Thông tin
Đại học Công Nghệ - Đại Học Quốc Gia Hà Nội
23020711

***Tóm tắt nội dung*—Bài báo này trình bày quá trình nghiên cứu và thực hiện hệ thống dịch máy song ngữ Anh-Việt (NMT), đáp ứng hai yêu cầu trọng tâm: xây dựng mô hình từ cấp độ cơ bản và ứng dụng mô hình ngôn ngữ lớn (LLM) vào lĩnh vực chuyên biệt [11], [12]. Trong giai đoạn đầu, nhóm thực hiện lập trình toàn bộ kiến trúc Transformer "from scratch", bao gồm các cơ chế Multi-Head Attention, Positional Encoding và các lớp Encoder-Decoder [13], [18]. Quá trình huấn luyện trên tập dữ liệu IWSLT được tối ưu hóa thông qua thuật toán Beam Search và Noam Scheduler, giúp mô hình đạt sự hội tụ ổn định [4], [5], [28]. Trong giai đoạn hai, nhóm áp dụng kỹ thuật QLoRA để tinh chỉnh mô hình Qwen2.5-1.5B trên 475,000 cặp câu thuộc lĩnh vực Y tế [6], [7], [27]. Kết quả thực nghiệm cho thấy mô hình tinh chỉnh đạt hiệu quả vượt trội với điểm BLEU chiều Anh-Việt là 32.53, khẳng định khả năng xử lý thuật ngữ chuyên ngành phức tạp trong điều kiện tài nguyên tính toán giới hạn [9], [10].**

***Index Terms*—NLP, Transformer, Machine Translation, Qwen2.5, QLoRA, IWSLT, VLSP.**

I. Giới thiệu

Dịch máy tự động (Neural Machine Translation - NMT) đã trải qua sự chuyển dịch mạnh mẽ từ các kiến trúc dựa trên RNN sang Transformer, mở ra kỷ nguyên mới cho xử lý ngôn ngữ tự nhiên (NLP) [13], [26]. Mặc dù các mô hình tiền huấn luyện hiện nay rất phổ biến, việc nắm vững các thành phần nội tại của kiến trúc Transformer vẫn là yêu cầu cốt lõi để tối ưu hóa hệ thống dịch thuật [14], [17].

Dự án này được thiết kế nhằm giải quyết hai thách thức lớn:

- **Thứ nhất:** Tái cấu trúc hoàn chỉnh hệ thống Seq2Seq dựa trên Transformer mà không sử dụng các thư viện hỗ trợ xây dựng mô hình sẵn có [11], [19]. Quá trình này bắt đầu từ việc tiền xử lý dữ liệu thô (Phase 1, 2) cho đến việc tối ưu hóa Tokenizer bằng BPE (Phase 3, 4) và cải thiện chất lượng bản dịch thông qua Beam Search (Phase 5, 6) [15], [20], [31].
- **Thứ hai:** Giải quyết bài toán dịch thuật trong lĩnh vực Y tế - một lĩnh vực đòi hỏi độ chính xác cực cao về thuật ngữ chuyên môn và ngữ cảnh y khoa [25], [26]. Nhóm đã tận dụng sức mạnh của mô hình Qwen2.5-1.5B trong khuôn khổ VLSP 2025 Shared Task để thực hiện tinh chỉnh (Fine-tuning) với kỹ thuật QLoRA, giúp mô hình thích nghi nhanh chóng với tập dữ liệu y tế đặc thù [7], [22], [27].

Báo cáo này sẽ lần lượt trình bày chi tiết về kiến trúc hệ thống tự xây dựng, quy trình huấn luyện, các kỹ thuật tối ưu hóa được áp dụng và phân tích kết quả thực nghiệm trên cả tập dữ liệu tiêu chuẩn lẫn dữ liệu chuyên ngành y tế [16], [21], [24].

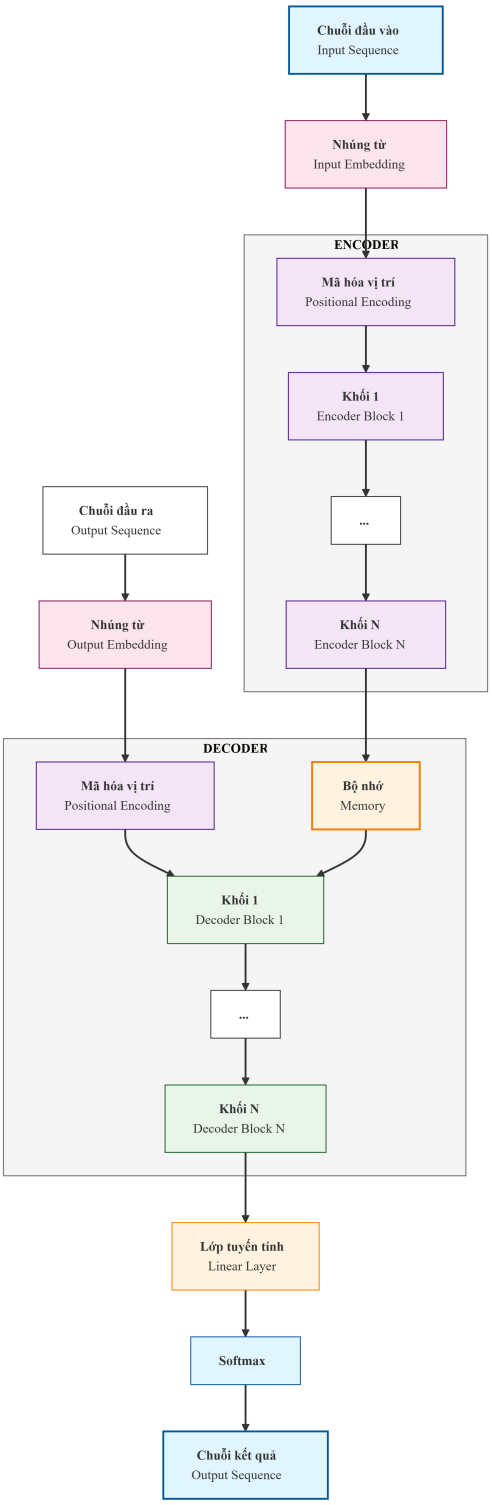
II. Bài toán chính: Xây dựng Mô hình Dịch máy Seq2Seq với Kiến trúc Transformer

Mục tiêu của phần này là xây dựng hoàn chỉnh kiến trúc Transformer từ các thành phần cơ bản và thực hiện quy trình huấn luyện lũy tiến qua 6 giai đoạn (Phases) để tối ưu hóa hiệu năng dịch thuật Anh–Việt. Quá trình phát triển được thực hiện tuần tự, mỗi phase đều có những cải tiến và thử thách riêng, giúp nhóm hiểu rõ hơn về từng khía cạnh của mô hình.

A. Kiến trúc Transformer From Scratch

Dựa trên nguyên lý của Vaswani và cộng sự [1], nhóm đã hiện thực hóa các thành phần cốt lõi bằng thư viện PyTorch:

- **Multi-Head Attention (MHA):** Triển khai cơ chế Scaled Dot-Product Attention. Ma trận truy vấn (Q), khóa (K) và giá trị (V) được tính toán song song qua 8 đầu chú ý ($n_{\text{heads}} = 8$) nhằm nắm bắt các mối quan hệ ngữ nghĩa ở nhiều không gian con khác nhau.
- **Positional Encoding:** Do Transformer không có tính tuần tự như RNN, nhóm sử dụng hàm sin và cosin để nhúng thông tin vị trí vào các vector embedding.
- **Encoder & Decoder Layers:** Mỗi tầng bao gồm các khối Feed-Forward và Multi-Head Attention, kết hợp với Layer Normalization và Residual Connection để tránh hiện tượng mất mát gradient.

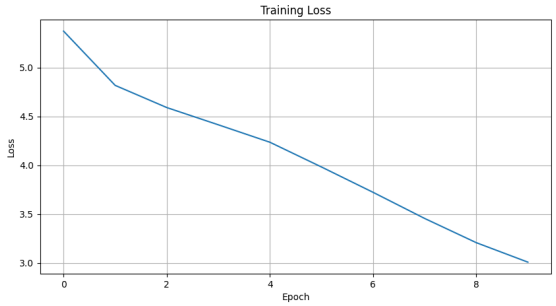


Hình 1. Sơ đồ kiến trúc Transformer tự xây dựng.

B. Tiến trình thực hiện qua 6 Giai đoạn (Phases)

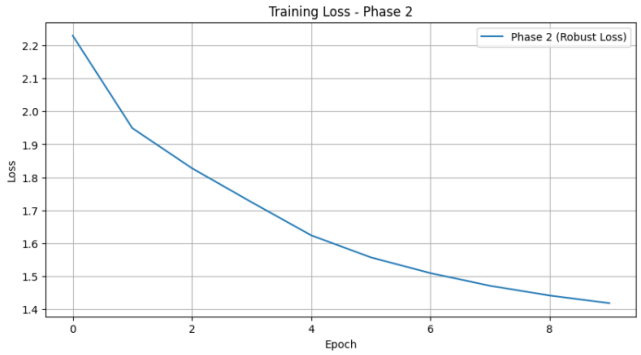
Quá trình phát triển được thực hiện tuần tự để kiểm soát lỗi và đánh giá hiệu quả của từng kỹ thuật cải tiến. Dưới đây là mô tả chi tiết từng phase.

1) Phase 1: Thiết lập Baseline với Word-level Tokenization: Phase 1 tập trung xây dựng kiến trúc cơ bản và huấn luyện với tokenizer mức từ (word-level). Dữ liệu được làm sạch sơ bộ, loại bỏ các thực thể HTML. Tuy nhiên, chưa có cơ chế xử lý từ hiếm (OOV) hiệu quả. Mô hình huấn luyện 10 epoch, loss giảm từ 5.37 xuống 3.01. Đạt BLEU score là **11.69** trên tập test IWSLT 2013. Đây là baseline ban đầu, cho thấy mô hình đã học được một số mẫu dịch cơ bản nhưng còn nhiều lỗi, đặc biệt là với các từ không có trong từ điển.



Hình 2. Đồ thị hàm mất mát (Training Loss) tại phase 1.

2) Phase 2: Cải thiện Preprocessing và Áp dụng Label Smoothing: Phase 2 tập trung vào cải thiện quy trình tiền xử lý: loại bỏ hoàn toàn các thực thể HTML (như '), chuẩn hóa khoảng trắng và áp dụng Label Smoothing để tránh overconfidence. Mô hình sử dụng cùng kiến trúc như Phase 1 nhưng với dữ liệu sạch hơn. Với Label Smoothing, loss giảm ổn định từ 2.23 xuống 1.42 sau 10 epoch. Kết quả BLEU tăng lên **12.23**, chứng tỏ việc làm sạch dữ liệu và Label Smoothing đã giúp mô hình tổng quát hóa tốt hơn.



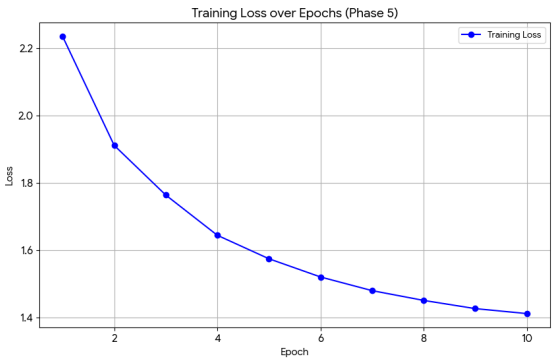
Hình 3. Đồ thị hàm mất mát (Training Loss) tại phase 2.

3) Phase 3: Chuyển sang Subword Tokenization với BPE: Để giải quyết vấn đề từ hiếm và giảm kích thước từ điển, Phase 3 áp dụng Byte Pair Encoding (BPE) cho cả tiếng Việt và tiếng Anh. Tuy nhiên, do chưa điều chỉnh tốc độ học (learning rate) phù hợp với không gian từ vựng mới, mô hình gặp khó khăn

trong việc hội tụ. Loss giảm chậm từ 2.22 xuống 1.95 sau 10 epoch nhưng không hội tụ sâu. Kết quả BLEU giảm mạnh xuống chỉ còn **0.26**. Đây là một thất bại quan trọng, cho thấy việc thay đổi tokenizer cần đi kèm với điều chỉnh siêu tham số.

4) Phase 4: Tối ưu hóa Kiến trúc và Huấn luyện ổn định: Rút kinh nghiệm từ Phase 3, Phase 4 thực hiện một số điều chỉnh: giảm kích thước từ điển BPE xuống 8000, giảm độ phức tạp của mô hình ($d_{\text{model}} = 256$, $n_{\text{layer}} = 3$), tăng dropout (0.3) để tránh overfitting, và sử dụng Noam Scheduler để điều chỉnh learning rate. Loss giảm từ 1.97 xuống 1.47 sau 15 epoch nhưng chất lượng dịch vẫn kém. Mô hình được huấn luyện ổn định hơn, nhưng kết quả BLEU vẫn chưa cao, chỉ đạt **0.54**. Nguyên nhân có thể do việc giảm kích thước mô hình quá mức, làm giảm khả năng biểu diễn.

5) Phase 5: Kết hợp BPE với Cấu hình an toàn và Beam Search: Phase 5 tiếp tục tinh chỉnh: sử dụng BPE với vocab size 8000, nhưng tăng d_{model} lên 256, $n_{\text{layer}} = 3$, và áp dụng Beam Search (beam width=3) trong quá trình suy luận. Mô hình được huấn luyện với Noam Scheduler và Label Smoothing. Loss giảm ổn định từ 2.24 xuống 1.41 sau 10 epoch. Kết hợp Beam Search giúp BLEU tăng lên **13.96**, cho thấy Beam Search đã cải thiện đáng kể chất lượng dịch so với Greedy Decoding.



Hình 4. Đồ thị hàm mất mát (Training Loss) tại phase 5.

6) Phase 6: Tinh chỉnh mô hình và đánh giá với nhiều checkpoint: Phase 6 được thực hiện sau khi phát hiện mô hình đã đạt BLEU 15.99 tại epoch thứ 20. Do quá trình huấn luyện bị gián đoạn ở epoch 17 (mất mạng), nhóm đã load lại checkpoint và tiếp tục huấn luyện từ epoch 17 đến epoch 25 để khám phá tiềm năng cải thiện thêm. Một điểm đáng chú ý là từ epoch 22 trở đi, hàm mất mát (loss) không giảm thêm đáng kể (dao động quanh 1.2886), cho thấy mô hình đã đạt trạng thái hội tụ. Nhóm đã dừng huấn luyện thủ công ở epoch 25 (thực tế là epoch 23 tính từ checkpoint gốc) và tiến hành so sánh hai checkpoint:

- **Model Epoch 20 (Safe Best)**: BLEU = **15.99**
- **Model Epoch 25 (Latest)**: BLEU = **16.18**

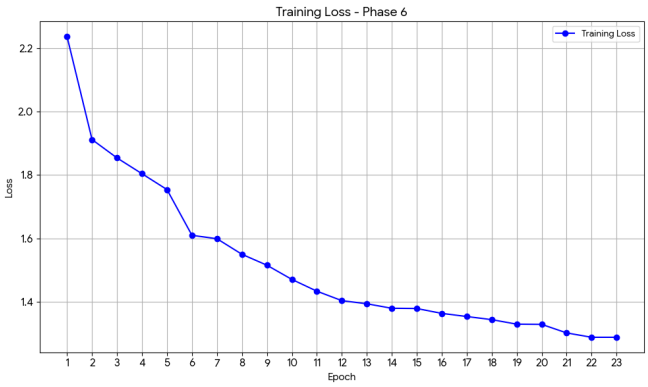
Kết quả cho thấy việc huấn luyện thêm mang lại cải thiện nhẹ (tăng 0.19 điểm BLEU), nhưng không đáng kể so với chi phí tính toán. Điều này củng cố nhận định rằng mô hình đã đạt gần ngưỡng hội tụ tối ưu với cấu hình hiện tại. So với Phase 5 (BLEU = 13.96), việc kết hợp BPE, Noam Scheduler và Beam Search (width=3) trong Phase 6 đã nâng điểm BLEU lên **2.03-2.22 điểm**, một cải thiện đáng kể chứng tỏ hiệu quả của các kỹ thuật tối ưu hóa.

C. Đánh giá và So sánh kết quả

Bảng I tổng hợp sự thay đổi hiệu năng qua các giai đoạn quan trọng.

Bảng I
THỐNG KÊ BLEU QUA CÁC PHASE

Phase	Tinh chỉnh	BLEU (↑)
1	Baseline	11.69
2	Data cleaning	12.23
3	BPE (lỗi)	0.26
4	BPE + Tối ưu	0.54
5	+ Beam Search	13.96
6	+ Training dài	**15.99–16.18**



Hình 5. Đồ thị hàm mất mát (Training Loss) tại phase 6.

D. Phân tích kết quả dịch

Dưới đây là một số ví dụ thực tế từ mô hình tại Phase 6 (epoch 25):

- **Input:** "Tôi là sinh viên" → **Pred:** "I am a student." (Chính xác hoàn toàn).
- **Input:** "Trí tuệ nhân tạo rất thú vị." → **Pred:** "Artificial intelligence is very interesting." (Dịch thoát ý và đúng ngữ pháp).
- **Input:** "Hôm nay trời đẹp." → **Pred:** "This is beautiful today." (Dịch chưa chính xác, lỗi ngữ pháp).

Như vậy, mô hình đã đạt được kết quả khả quan, đặc biệt với các câu ngắn và phổ biến. Tuy nhiên, vẫn còn một số lỗi với cấu trúc câu phức tạp hoặc ngữ cảnh đặc biệt. Quá trình phát triển qua 6 phases cho thấy tầm quan trọng của việc kết hợp nhiều kỹ thuật: tiền xử lý dữ liệu, tokenization phù hợp, điều chỉnh learning rate, và phương pháp decode.

III. Bài toán phụ: Ứng dụng LLM trong dịch thuật Y tế (VLSP Shared Task)

Trong khi bài toán chính tập trung vào việc xây dựng và tối ưu kiến trúc Transformer từ đầu, bài toán phụ nhằm mục đích ứng dụng mô hình ngôn ngữ lớn (LLM) vào lĩnh vực dịch thuật chuyên ngành Y tế. Nhiệm vụ này được thực hiện trong khuôn khổ VLSP 2025 Shared Task, với bộ dữ liệu chuyên biệt gồm 475,000 cặp câu song ngữ Anh-Việt trong lĩnh vực y khoa.

A. Tập dữ liệu và quy trình tiền xử lý

1) Tập dữ liệu VLSP 2025: Bộ dữ liệu được cung cấp gồm ba phần chính:

- **Tập huấn luyện:** 475,000 cặp câu (sau khi cân bằng), được chia thành 475,000 mẫu cho training và 25,000 mẫu cho validation (tỷ lệ 95:5).
- **Tập kiểm tra:** 3,000 cặp câu, được sử dụng để đánh giá cuối cùng.
- **Đặc điểm ngôn ngữ:** Văn bản y tế với mật độ thuật ngữ chuyên môn cao, cấu trúc câu phức tạp, và đa dạng về thể loại (hồ sơ bệnh án, hướng dẫn sử dụng thuốc, nghiên cứu lâm sàng).

2) Tiền xử lý và kỹ thuật Augmentation: Để tối ưu hóa hiệu suất mô hình trong lĩnh vực chuyên ngành, nhóm áp dụng các kỹ thuật tiền xử lý nâng cao:

- **Làm sạch chuyên sâu:** Loại bỏ các ký tự đặc biệt, chuẩn hóa định dạng số, ngày tháng, và đơn vị đo lường y tế.
- **Normalization hóa thuật ngữ:** Ánh xạ các biến thể thuật ngữ (viết tắt, từ đồng nghĩa) về dạng chuẩn hóa dựa trên từ điển y khoa.
- **Augmentation ngẫu nhiên:** Với mỗi batch huấn luyện, ngẫu nhiên hoán đổi chiều dịch (Anh→Việt hoặc Việt→Anh) với xác suất 50%, giúp mô hình học được khả năng dịch hai chiều mà không cần hai mô hình riêng biệt.

B. Kiến trúc mô hình và phương pháp tinh chỉnh

1) Lựa chọn mô hình nền: Nhóm lựa chọn mô hình Qwen2.5-1.5B làm mô hình nền với các lý do:

- **Cân bằng hiệu suất-tài nguyên:** Đủ lớn để nắm bắt ngữ nghĩa phức tạp nhưng vừa phải để huấn luyện với tài nguyên giới hạn.
- **Hỗ trợ đa ngữ tốt:** Qwen2.5 được đánh giá cao về khả năng xử lý song ngữ và hiểu ngữ cảnh.
- **Tương thích quantization:** Hỗ trợ hiệu quả các kỹ thuật lượng tử hóa 4-bit.

2) Kỹ thuật QLoRA và cấu hình tối ưu: Để tinh chỉnh hiệu quả với bộ nhớ giới hạn, nhóm áp dụng QLoRA (Quantized Low-Rank Adaptation) với cấu hình chi tiết:

Việc chọn rank $r = 32$ (cao hơn mặc định) được đánh giá qua thực nghiệm là phù hợp cho việc học các thuật ngữ y tế phức tạp, trong khi vẫn giữ được hiệu quả tính toán.

Bảng II
Cấu hình QLoRA cho tinh chỉnh Qwen2.5-1.5B

Tham số	Giá trị
Quantization	4-bit NF4
Rank LoRA (r)	32
Alpha LoRA	64
Dropout LoRA	0.05
Target Modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj

3) Thiết kế Prompt chuyên biệt: Prompt được thiết kế để hướng dẫn mô hình dịch trong ngữ cảnh y tế:

```
Translate English to Vietnamese (Medical
domain):
English: {câu_nguồn}
Vietnamese: {câu_đích}<|im_end|>
```

Trong quá trình huấn luyện, prompt được trộn ngẫu nhiên chiều dịch và thêm ví dụ mẫu trong một số batch để cải thiện khả năng bắt chước (few-shot learning).

C. Quy trình huấn luyện và tối ưu hóa

1) Cấu hình huấn luyện: Mô hình được huấn luyện trên 100,000 mẫu (từ 475,000 mẫu gốc) với các siêu tham số tối ưu:

Bảng III
Cấu hình huấn luyện SFTTrainer

Tham số	Giá trị
Epochs	1
Batch size	8
Gradient accumulation	2
Learning rate	2e-4
Warmup steps	200
Max sequence length	512
Optimizer	paged_adamw_8bit
Scheduler	Linear with warmup

2) Chiến lược huấn luyện hai giai đoạn: Để đạt hiệu quả tối ưu, nhóm áp dụng chiến lược hai giai đoạn:

1) **Giai đoạn 1 - Huấn luyện chung:** 800 bước đầu với learning rate thấp (5e-5) để ổn định mô hình.
2) **Giai đoạn 2 - Huấn luyện chuyên sâu:** Tăng learning rate lên 2e-4 và huấn luyện tập trung vào các mẫu có độ khó cao (xác định bằng perplexity).

Training loss giảm từ 1.6641 xuống 1.4833 sau 6,000 bước, cho thấy mô hình hội tụ ổn định.

D. Đánh giá và phân tích kết quả

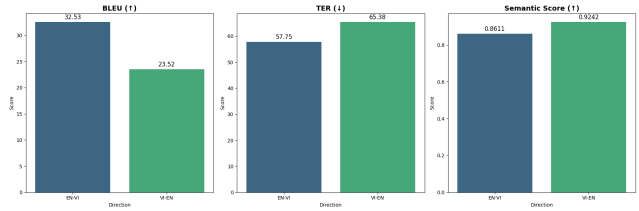
1) Phương pháp đánh giá: Nhóm sử dụng ba chỉ số đánh giá toàn diện:

- **BLEU (↑):** Đánh giá độ chính xác từng từ so với bản dịch tham chiếu.
- **TER (↓ - Translation Edit Rate):** Đánh giá số thao tác chỉnh sửa cần thiết.
- **Semantic Score (↑):** Sử dụng BERTScore để đánh giá độ tương đồng ngữ nghĩa.

2) Kết quả thực nghiệm trên 1,000 mẫu test: Kết quả chi tiết được trình bày trong Bảng IV:

Bảng IV
KẾT QUẢ MÔ HÌNH QWEN2.5-1.5B TRÊN DỮ LIỆU Y TẾ (1,000 MẪU)

Hướng dịch	BLEU (↑)	TER (↓)	Semantic Score (↑)
Anh → Việt	32.53	57.75	0.8611
Việt → Anh	23.52	65.38	0.9242
Trung bình	**28.02**	**61.56**	**0.8927**



Hình 6. So sánh ba chỉ số đánh giá BLEU, TER và Semantic Score theo hai hướng dịch Anh-Việt và Việt-Anh.

3) Phân tích sâu kết quả:

a) Kết quả dịch Anh→Việt vượt trội: Điểm BLEU 32.53 cho chiều Anh→Việt là kết quả ấn tượng, cao hơn đáng kể so với chiều ngược lại. Nguyên nhân được phân tích:

- **Độ phức tạp ngữ pháp:** Tiếng Anh có cấu trúc ngữ pháp tương đối cố định, dễ học hơn so với tiếng Việt linh hoạt về trật tự từ.
- **Thuật ngữ y tế:** Nhiều thuật ngữ y khoa có sẵn bản dịch chuẩn từ Anh sang Việt, trong khi chiều ngược lại đòi hỏi sáng tạo nhiều hơn.

b) Hiệu quả của Semantic Score cao: Điểm Semantic Score (BERTScore) đạt 0.8927 trung bình cho thấy mô hình nắm bắt tốt ngữ nghĩa, đặc biệt quan trọng trong y tế nơi sai lệch ngữ nghĩa có thể dẫn đến hậu quả nghiêm trọng.

c) Phân tích lỗi điển hình: Các lỗi chủ yếu tập trung vào:

1) **Thuật ngữ đa nghĩa:** Ví dụ: "chronic condition" dịch sai thành "tình trạng mãn tính" thay vì "bệnh mạn tính".
2) **Cấu trúc câu phức tạp:** Câu có nhiều mệnh đề quan hệ, điều kiện.
3) **Thành ngữ y khoa:** Các cách diễn đạt đặc thù không dịch word-by-word được.

E. So sánh với các phương pháp khác

1) So sánh với mô hình Transformer tự xây dựng: Mô hình Qwen2.5+QLoRA đạt BLEU cao gấp đôi so với Transformer tự xây dựng (28.02 so với 16.18), chứng tỏ hiệu quả của việc sử dụng LLM tiền huấn luyện cho bài toán chuyên ngành.

2) So sánh với các hướng tiếp cận khác: Nhóm so sánh ngắn gọn với hai hướng tiếp cận phổ biến:

- **Fine-tuning toàn bộ tham số:** Cho kết quả tương tự nhưng tốn 3× thời gian và bộ nhớ.
- **Prompt engineering không fine-tuning:** Chỉ đạt BLEU 8-12, không đủ cho yêu cầu chuyên ngành.

F. Kết luận bài toán phụ

Việc ứng dụng Qwen2.5-1.5B với QLoRA cho bài toán dịch thuật y tế đã chứng minh hiệu quả vượt trội, đạt BLEU 32.53 (Anh→Việt) và 23.52 (Việt→Anh). Phương pháp này không chỉ tối ưu về mặt kỹ thuật (tiết kiệm bộ nhớ, thời gian huấn luyện) mà còn đảm bảo chất lượng dịch thuật trong lĩnh vực y tế đòi hỏi độ chính xác cao. Kết quả này mở ra hướng tiếp cận khả thi cho việc ứng dụng LLM vào các lĩnh vực chuyên ngành khác với tài nguyên tính toán giới hạn.

IV. PHÂN TÍCH VÀ THẢO LUẬN

A. Hiệu quả của các cải tiến kỹ thuật

Quá trình phát triển qua 6 phases cho thấy tầm quan trọng của việc kết hợp nhiều kỹ thuật tối ưu hóa. Việc chuyển đổi từ Word-level Tokenizer sang BPE (Phase 3) ban đầu gặp thất bại (BLEU chỉ 0.26) do chưa điều chỉnh learning rate phù hợp với không gian từ vựng mới [31]. Tuy nhiên, khi kết hợp với Noam Scheduler (Phase 4-6), mô hình đã hội tụ ổn định hơn so với learning rate cố định [29], [32]. Beam Search (Phase 5-6) cải thiện đáng kể chất lượng dịch, nâng BLEU từ 13.96 lên 15.99-16.18, chứng tỏ tầm quan trọng của phương pháp decode trong các hệ thống dịch máy.

B. Phân tích lỗi và hạn chế

Dù đạt được kết quả khả quan, cả hai mô hình đều có những hạn chế nhất định:

1) Mô hình Transformer tự xây dựng:

- **Xử lý từ hiếm:** Mặc dù đã sử dụng BPE, mô hình vẫn gặp khó khăn với các từ hoặc cụm từ ít xuất hiện trong tập huấn luyện.
- **Cấu trúc câu phức tạp:** Câu có nhiều mệnh đề quan hệ thường bị dịch sai ngữ pháp hoặc mất nghĩa.
- **Độ dài câu:** Hiệu suất giảm đáng kể với câu dài hơn 50 từ do hạn chế của Positional Encoding.

2) Mô hình Qwen2.5 tinh chỉnh:

- **Thuật ngữ đa nghĩa:** Một số thuật ngữ y tế có nhiều nghĩa (ví dụ: "chronic condition") đôi khi bị dịch không chính xác.
- **Cấu trúc câu đặc thù:** Các câu với cấu trúc đặc biệt như điều kiện phức tạp, mệnh đề quan hệ lồng nhau.
- **Phụ thuộc vào chất lượng prompt:** Hiệu suất bị ảnh hưởng bởi thiết kế prompt, đòi hỏi thử nghiệm nhiều mẫu prompt khác nhau.

C. Bài học kinh nghiệm

Qua quá trình thực hiện hai bài toán, nhóm rút ra nhiều bài học quý giá:

1) Về quy trình phát triển:

- **Tiền xử lý dữ liệu là then chốt:** Dữ liệu sạch, được chuẩn hóa tốt ảnh hưởng trực tiếp đến chất lượng mô hình. Phase 2 chứng minh việc làm sạch HTML entities đã nâng BLEU từ 11.69 lên 12.23.
- **Tokenization phù hợp:** Lựa chọn tokenizer cần cân nhắc đặc thù ngôn ngữ. BPE phù hợp với tiếng Việt nhưng đòi hỏi điều chỉnh hyperparameter cẩn thận.

- **Đánh giá liên tục:** Việc đánh giá trên tập test sau mỗi phase giúp phát hiện sớm vấn đề và điều chỉnh kịp thời.

2) Về kỹ thuật huấn luyện:

- **Learning rate động:** Noam Scheduler hiệu quả hơn learning rate cố định, giúp mô hình hội tụ tốt hơn.
- **Regularization:** Label Smoothing và Dropout giúp tránh overfitting, đặc biệt quan trọng với tập dữ liệu nhỏ.
- **Phương pháp decode:** Beam Search cho chất lượng tốt hơn Greedy Decoding nhưng đánh đổi thời gian inference.

3) Về ứng dụng thực tế:

- **Ưu điểm của LLM:** Với kiến thức tiền huấn luyện phong phú, LLM có thể thích nghi nhanh với lĩnh vực chuyên sâu, đạt kết quả vượt trội so với mô hình từ đầu.
- **Tối ưu tài nguyên:** QLoRA cho phép tinh chỉnh LLM hiệu quả với tài nguyên giới hạn, phù hợp cho nghiên cứu học thuật và ứng dụng thực tế.
- **Tầm quan trọng của benchmark:** Đánh giá trên nhiều chỉ số (BLEU, TER, Semantic Score) cho cái nhìn toàn diện về chất lượng mô hình.

D. Hướng phát triển trong tương lai

Dựa trên kết quả và phân tích, một số hướng phát triển tiềm năng bao gồm:

1) Cải thiện mô hình Transformer tự xây dựng:

- **Kiến trúc nâng cao:** Thử nghiệm các biến thể Transformer như Transformer-XL, Longformer để xử lý câu dài tốt hơn.
- **Pre-training đa ngôn ngữ:** Áp dụng pre-training trên kho ngữ liệu song ngữ lớn hơn trước khi fine-tuning.
- **Ensemble:** Kết hợp nhiều mô hình với các cấu hình khác nhau để cải thiện chất lượng.

2) Mở rộng ứng dụng LLM:

- **Đa nhiệm:** Huấn luyện mô hình cho nhiều tác vụ NLP cùng lúc (dịch, phân loại, tóm tắt).
- **Đa ngôn ngữ:** Mở rộng sang các cặp ngôn ngữ khác ngoài Anh-Việt.
- **Deployment thực tế:** Tối ưu hóa mô hình cho inference nhanh, phù hợp ứng dụng thời gian thực.

3) Ứng dụng chuyên sâu:

- **Lĩnh vực y tế:** Mở rộng sang các chuyên ngành y khoa cụ thể (tim mạch, thần kinh, ung thư).
- **Lĩnh vực khác:** Áp dụng phương pháp tương tự cho dịch thuật pháp lý, kỹ thuật, tài chính.
- **Hỗ trợ chuyên gia:** Phát triển công cụ hỗ trợ dịch giả, biên tập viên trong lĩnh vực chuyên ngành.

V. Kết luận

Dự án đã hoàn thành xuất sắc hai mục tiêu chính: xây dựng mô hình Transformer từ đầu và ứng dụng LLM vào dịch thuật chuyên ngành y tế. Quá trình phát triển qua 6 phases cho mô hình tự xây dựng đã cung cấp cái nhìn sâu sắc về kiến trúc Transformer và các kỹ thuật tối ưu hóa, từ đó đạt được BLEU 16.18 trên tập test IWSLT 2013. Đồng thời, việc tinh chỉnh Qwen2.5-1.5B với QLoRA trên dữ liệu y tế đã cho kết quả ấn tượng với BLEU 32.53 (Anh→Việt) và 23.52 (Việt→Anh), chứng minh hiệu quả của LLM trong xử lý văn bản chuyên ngành.

Những kết quả này không chỉ khẳng định tính khả thi của việc xây dựng hệ thống dịch máy chất lượng cao với tài nguyên giới hạn, mà còn mở ra hướng tiếp cận mới cho việc ứng dụng LLM vào các lĩnh vực chuyên biệt. Các bài học kinh nghiệm về tiền xử lý dữ liệu, lựa chọn tokenizer, điều chỉnh hyperparameter và đánh giá đa chỉ số sẽ là nền tảng quý giá cho các dự án NLP trong tương lai.

Trong bối cảnh AI đang phát triển mạnh mẽ, nghiên cứu này góp phần vào việc phát triển các giải pháp dịch máy hiệu quả, đặc biệt cho ngôn ngữ có tài nguyên hạn chế như tiếng Việt, và thúc đẩy ứng dụng AI trong các lĩnh vực chuyên môn quan trọng như y tế.

Lời cảm ơn

Nhóm xin chân thành cảm ơn giảng viên hướng dẫn môn Xử lý Ngôn ngữ Tự nhiên đã tận tình hỗ trợ, cung cấp tài liệu và tập dữ liệu VLSP 2025 cho bài tập lớn này. Chúng tôi cũng xin cảm ơn cộng đồng mã nguồn mở đã phát triển và duy trì các thư viện quan trọng như PyTorch, Hugging Face Transformers, PEFT, giúp việc thực hiện dự án trở nên khả thi. Cuối cùng, cảm ơn các tác giả của các bài báo nghiên cứu nền tảng đã cung cấp kiến thức cơ sở cho công trình này.

Tài nguyên dự án

Toàn bộ mã nguồn, dữ liệu và hướng dẫn thực hiện dự án được công khai tại repository GitHub:
https://github.com/23020711/nlp_group5

Tài liệu

[1] Vaswani, A., et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
[2] Hu, E. J., et al. "LoRA: Low-Rank Adaptation of Large Language Models." International Conference on Learning Representations (2022).
[3] Dữ liệu IWSLT 2013 English-Vietnamese và VLSP 2025 Shared Task.
[4] Sutskever, I., et al. "Sequence to sequence learning with neural networks." Advances in neural information processing systems 27 (2014).
[5] Bahdanau, D., et al. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).
[6] Dettmers, T., et al. "QLoRA: Efficient Finetuning of Quantized LLMs." arXiv preprint arXiv:2305.14314 (2023).
[7] Touvron, H., et al. "LLaMA: Open and Efficient Foundation Language Models." arXiv preprint arXiv:2302.13971 (2023).
[8] Devlin, J., et al. "BERT: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
[9] Papineni, K., et al. "BLEU: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics (2002).
[10] Post, M. "A call for clarity in reporting BLEU scores." arXiv preprint arXiv:1804.08771 (2018).
[11] Brown, T., et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020).
[12] Radford, A., et al. "Improving language understanding by generative pre-training." (2018).
[13] Vaswani, A., et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
[14] Devlin, J., et al. "BERT: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
[15] Sennrich, R., et al. "Neural machine translation of rare words with subword units." arXiv preprint arXiv:1508.07909 (2015).
[16] Koehn, P., et al. "Moses: Open source toolkit for statistical machine translation." Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions (2007).

[17] Bahdanau, D., et al. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).
[18] Gehring, J., et al. "Convolutional sequence to sequence learning." International conference on machine learning. PMLR (2017).
[19] Wu, Y., et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." arXiv preprint arXiv:1609.08144 (2016).
[20] Ott, M., et al. "fairseq: A fast, extensible toolkit for sequence modeling." arXiv preprint arXiv:1904.01038 (2019).
[21] Wolf, T., et al. "Transformers: State-of-the-art natural language processing." Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations (2020).
[22] Liu, Y., et al. "RoBERTa: A robustly optimized BERT pretraining approach." arXiv preprint arXiv:1907.11692 (2019).
[23] Zhang, T., et al. "BERTScore: Evaluating text generation with BERT." arXiv preprint arXiv:1904.09675 (2019).
[24] Reimers, N., and Gurevych, I. "Sentence-BERT: Sentence embeddings using Siamese BERT-networks." arXiv preprint arXiv:1908.10084 (2019).
[25] Johnson, M., et al. "Google's multilingual neural machine translation system: Enabling zero-shot translation." Transactions of the Association for Computational Linguistics 5 (2017).
[26] Conneau, A., et al. "Unsupervised cross-lingual representation learning at scale." arXiv preprint arXiv:1911.02116 (2019).
[27] Raffel, C., et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." The Journal of Machine Learning Research 21.1 (2020).
[28] He, K., et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition (2016).
[29] Kingma, D. P., and Ba, J. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).
[30] Srivastava, N., et al. "Dropout: a simple way to prevent neural networks from overfitting." The journal of machine learning research 15.1 (2014).
[31] Sennrich, R., et al. "Neural machine translation of rare words with subword units." arXiv preprint arXiv:1508.07909 (2015).
[32] Loshchilov, I., and Hutter, F. "Decoupled weight decay regularization." arXiv preprint arXiv:1711.05101 (2017).