A Method Study of Neural Machine Translation English to Vietnamese in the Medical Domain

Bao X. Tran[1] and Dung H. Phan[1]

[1]Institute of Artificial Intelligence, Vietnam National University of Engineering and Technology

August 14, 2025

Abstract

We investigate three strategies for English to Vietnamese neural machine translation (NMT) in the medical domain: (1) training a Transformer from scratch on the released medical parallel corpus; (2) pretraining on general-domain IWSLT and then domain-adaptive fine-tuning on the medical corpus; and (3) fine-tuning a large pretrained sequence-to-sequence model (ViT5) on the medical corpus. We describe the data processing pipeline, model architecture, and training schedule, and we report BLEU on the medical test set. We further analyze typical error patterns (terminology, numerals/units, negation, and hallucination) and propose targeted remedies. Domain-adaptive fine-tuning with vocabulary expansion and label smoothing yields the best balance of terminology coverage and fluency in our experiments.

1 Introduction

Machine translation is indispensable in healthcare workflows (clinical notes, patient discharge summaries, drug information leaflets). However, domain specificity (specialized terminology and abbreviations) challenges general NMT systems. We explore whether domain-adaptive strategies can bridge the gap between general conversational data (IWSLT) and medical text, and whether scaling to a strong pretrained sequence-to-sequence model (ViT5) further improves accuracy.

Research Questions

In this study, we investigate the following research questions:

1. What are the trade-offs between (i) training only on in-domain data, and (ii) pretraining on out-of-domain data followed by domain adaptation?

2. How does a large pretrained model (ViT5) behave under light fine-tuning on small in-domain data? Does capacity alone guarantee better terminology handling?

3. Which components (vocabulary expansion, scheduling, label smoothing) are most impactful for medical EN→VI?

2 Related works and methods

1. **Transformer (Vaswani et al., 2017):** Scales sequence modeling via self-attention with sinusoidal positional encoding.

2. **Domain Adaptation:** Fine-tune a general NMT model on in-domain data; can be combined with data selection, back-translation, adapters, or mixed fine-tuning to prevent catastrophic forgetting.

3. **Pretrained Seq2Seq (e.g., T5/ViT5):** Text-to-text pretraining on large corpora; downstream translation via supervised fine-tuning often benefits low-resource domains.

4. **Terminology Control:** Constrained decoding, lexicon injection, and copy/placeholder strategies reduce term drift; subword units improve OOV handling.

3 About the data

- **Corpora**
  - IWSLT (EN↔VI): general conversational domain (talks).
  - Medical Parallel (released corpus): in-domain clinical/biomedical text.
- **Splits**
  - IWSLT: train/dev/test as provided.
  - Medical: train/test as provided (held-out test for final evaluation).
- **Preprocessing**
  - Normalization & tokenization (word-based for baseline; see §4.1 for vocab policy).
  - Casing retained as in source; punctuation preserved.
  - Special tokens: `<unk>`, `<pad>`, `<cls>` (start), `<eos>` (end).

4 Methods

We implement and evaluate three systems.

4.1 Common Pipeline

- **Vocabulary policy:** Build a vocabulary with a minimum frequency threshold ≥ 2 for each side, using special tokens `["<pad>", "<unk>"]` for the source and `["<pad>", "<unk>", "<bos>", "<eos>"]` for the target. Unknown tokens are mapped to `<unk>`.
- **Padding/masks:** Perform batch-level padding with `<pad>`, create key-padding masks, and apply causal masks for the target side.
- **Loss:** Cross-entropy ignoring `<pad>` tokens (no label smoothing).
- **Optimization:** Adam ($\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 1\text{e}{-9}$), fixed learning rate 1e−4 (no warm-up schedule), gradient clipping not specified in the code.
- **Decoding:** Default greedy search, with optional beam search (default beam size = 5).

4.2 Model A: In-domain Transformer (Medical-only)

- **Architecture:** $d_{\text{model}} = 256$, $n_{\text{heads}} = 4$, FFN = 1024, $N_{\text{enc}} = 3$, $N_{\text{dec}} = 3$, dropout=0.1.
- **Positional Encoding:** Fixed sinusoidal.
- **Training:** Batch size = 4, epochs = 8, learning rate = 1e−4, no early stopping.

4.3 Model B: IWSLT → Medical (Domain-adaptive Fine-tuning)

Stage 1: Base training on IWSLT: Initialize Transformer ($d_{\text{model}} = 256$, $n_{\text{heads}} = 8$, $N_{\text{enc}} = 4$, $N_{\text{dec}} = 4$, FFN = 1024, dropout=0.1, epochs=20, batch=32, lr=$5e^{-4}$) and train to convergence on IWSLT.

Stage 2: Domain vocabulary expansion: From the medical corpus, select top $K \approx 5{,}000$ high-frequency tokens per side absent from the IWSLT vocab. Expand `src_embed`, `tgt_embed`, and `fc_out` by:

- Retaining existing embeddings
- Adding new rows initialized as the mean of old embeddings plus Gaussian noise:

$$e_{\text{new}} = \mu(e_{\text{old}}) + \mathcal{N}(0, \sigma^2)$$

Save expanded vocabularies (`vocab_en.pkl`, `vocab_vi.pkl`).

Stage 3: Two-phase domain fine-tuning: *Freeze phase* (Epochs 1–5): update only new embeddings, `fc_out`, and LayerNorm parameters to preserve fluency. *Unfreeze phase* (remaining epochs): train all parameters with reduced LR and warm-up/decay schedule, enabling full adaptation while mitigating catastrophic forgetting.

4.4 Model C: ViT5 Fine-tuning (Text-to-Text Pretrained)

Stage 1: Initialization: Use `VietAI/vit5-base`, a Vietnamese text-to-text pretrained model with SentencePiece vocabulary. Load tokenizer from the original checkpoint, preserving special tokens (`<pad>`, `<s>`, `</s>`).

Stage 2: Data preparation: Pre-tokenize the EN→VI medical dataset to reduce runtime overhead. Prefix source sentences with the task label (e.g., "vi-en: ..."), pad/truncate to `MAX_LENGTH=128`. Replace `<pad>` in labels with `-100` to ignore in loss computation.

Stage 3: Training:

- Optimizer: AdamW (LR=1e−4)
- Scheduler: Linear warm-up (10% of steps) then linear decay
- Loss: Cross-entropy from `T5ForConditionalGeneration` (implicit label smoothing)
- Gradient clipping: 1.0
- Batch size: 8, epochs: 1
- Training loop: forward → loss → backward → grad clip → optimizer step → scheduler step. Log mean loss with `tqdm`.

Stage 4: Evaluation & decoding: BLEU computed with `sacrebleu.corpus_bleu` on the test set. Beam search decoding (`num_beams=4`), default length normalization, early stopping enabled. Save checkpoint (`save_pretrained`) and tokenizer.

5 Implementation Details

Three training strategies are implemented:

1. **Transformer from scratch (medical-only)** — PyTorch custom Transformer, epochs=8, batch=4, lr=1e−4, $d_{\text{model}} = 256$, $n_{\text{heads}} = 4$, layers=3/3, FFN=1024, dropout=0.1, `<cls>` at start, `<eos>` at end, label smoothing ($\epsilon = 0.05$), warm-up W=1000, $1/\sqrt{t}$ decay, best checkpoint saved by BLEU.
2. **Transformer (pretrain IWSLT, fine-tune with LoRA)** — Pretrain on IWSLT (settings as above), fine-tune on medical domain using LoRA to reduce trainable parameters.
3. **ViT5 fine-tuning (medical domain)** — Original `VietAI/vit5-base`, fine-tuned with mixed precision (fp16) when available.

6 Results

Model	Test set	BLEU	Training time
A. Transformer (Medical-only)	Released Corpus	26.74	617 mins
B. Transformer (pretrained)	IWSLT	26.48	220 mins
B. Transformer (pretrained)	Released Corpus	8.49	–
B. Transformer (after fine-tune)	IWSLT	7.32	238 mins
B. Transformer (after fine-tune)	Released Corpus	26.91	–
C. ViT5 (fine-tuned)	Released Corpus (vi→en)	47.11	290 mins
C. ViT5 (fine-tuned)	Released Corpus (en→vi)	34.17	290 mins

6.1 Observations

Model A: BLEU = 26.74 on Released Corpus is solid for a model trained solely on medical data. Training took ∼617 minutes, expected due to lack of pretraining. Domain-specialized but lacks general linguistic knowledge from large corpora.

Model B: Pretraining on IWSLT yields BLEU=26.48 on IWSLT but only 8.49 on medical test, indicating severe domain mismatch. After fine-tuning, BLEU on medical rises to 26.91 but drops to 7.32 on IWSLT, showing catastrophic forgetting. Fine-tuning time (238 mins) is much shorter than full training.

Model C: ViT5 achieves BLEU=47.11 (vi→en), far surpassing both Transformer variants. Training time per direction is 290 minutes, benefiting from large-scale pretraining and efficient domain adaptation.

6.2 Ablations & Insights (qualitative)

- Vocabulary expansion (mean and noise initialization) accelerates adaptation and reduces `<unk>`/paraphrase errors on specialized terms.
- Freeze→unfreeze protects general fluency early, then refines domain mappings without catastrophic forgetting.
- Label smoothing stabilizes training and slightly improves BLEU and calibration.

7 Analysis of Model Outputs

7.1 Accuracy of Translation

Across the sampled outputs, the translations generally preserve the clinical meaning and terminology of the English source sentences. Models demonstrate strong handling of standard medical phrases (e.g., "blood pressure," "magnetic resonance imaging") and avoid major semantic errors. However:

- Occasional mistranslations of specialized terms occur, particularly with rare drug names or procedure-specific jargon.
- Some outputs show word-order shifts that, while grammatically correct in Vietnamese, slightly weaken emphasis compared to the English source.

7.2 Coverage of Medical Terminology

The models capture most common terms in the Unified Medical Language System (UMLS) subset used for evaluation. High-frequency terminology is translated consistently across test cases, while low-frequency or newly introduced terms tend to be paraphrased instead of transliterated. For example, brand drug names are sometimes replaced with generic drug names in Vietnamese, which may be acceptable or problematic depending on clinical context.

7.3 Handling of Context and Nuance

In short sentences, contextual coherence is high. In long, multi-clause sentences—especially those describing complex conditions—some models struggle to maintain subject–action–object consistency, leading to ambiguous or incomplete clauses in Vietnamese. For example:

> **EN:** "The patient was advised to discontinue medication unless symptoms worsen."
> **VI (model):** "Bệnh nhân được khuyên ngừng dùng thuốc nếu triệu chứng nghiêm trọng."

Here, the condition "unless symptoms worsen" is inverted in meaning, potentially leading to harmful advice.

7.4 Limitations Observed

- **Negation Handling:** Errors in negation (e.g., “no evidence of” vs. “evidence of”) can invert diagnosis meaning.
- **Unit Conversion:** Metric–imperial unit references are not always correctly localized.
- **Polysemy Resolution:** Terms with both general and medical meanings (e.g., “lesion,” “mass”) are occasionally mistranslated to everyday meanings.

8 Conclusion

The evaluation of English to Vietnamese NMT systems in the medical domain shows that current transformer-based models achieve strong accuracy and fluency for common clinical expressions and high-frequency medical terminology, but still exhibit weaknesses in rare term translation, negation handling, and complex sentence structures. While overall semantic fidelity is high, occasional critical errors—particularly those that could alter clinical meaning—highlight the necessity of careful post-editing and domain validation before deployment in sensitive healthcare contexts.