Report 1

1. Introduction

In this report, I will present my understanding of machine translation, its applications, and the techniques currently employed in this field. Furthermore, I evaluate several machine translation models mentioned in the paper *"Improving Vietnamese-English Medical Machine Translation"* using the MedEV dataset. The evaluation is conducted in both translation directions: Vietnamese-to-English and English-to-Vietnamese. The performance of each model is measured using the BLEU score, which will be reported and discussed in this document.

Due to the computational limitations of the current setup, only a small fraction of the test set is used for the evaluation. Finally, this report provides observations based on the obtained results and outlines the plan for the next phase of the study.

2. What is Machine translation?

Machine translation is the process of using Article Intelligence to automatically translate text from one language to another without human involvement.

Machine translation works following a basic 2 step process:

1 – Decode the source language meaning of the original text

2 – Encode the meaning into the target language

There are several approaches on how to implement the above process:

Rule base Machine translation (RBMT)

RBMT relies on manually defined linguistic rules and dictionaries, offering strong grammatical control but struggling with idiomatic expressions and language diversity.

Statical Machine translation (SMT)

SMT uses statistical models trained on large parallel corpora to identify translation patterns, improving flexibility but requiring extensive data and failing to capture long-range dependencies.

Neural Machine translation (NMT)

NMT, the current dominant approach, employs neural networks—especially transformer-based architectures—to learn translations end-to-end. It effectively captures contextual meaning, handles idioms, and produces fluent translations. While RBMT and SMT still have niche uses, NMT has largely replaced them due to its superior performance and contextual understanding.

3. Why NMT is well-suited for medical translation

There are several reasons explaining why NMT is better than RBMT and SMT in medical translation.

Superior performance

NMT models have significantly improved translation quality compared to earlier approaches such as rule-based machine translation (RBMT) and statistical machine translation (SMT). NMT model shave the ability to capture long-range dependencies, handle language variations, and produce more fluent and natural-sounding translations. This property is attributed to the inherent sequential nature of neural networks, enabling them to model context and relationships more effectively.

Context-aware translation

The key advantages of NMT models, especially transformer-based architectures, is their ability to capture contextual information. Transformers utilize self-attention mechanisms, allowing the model to focus on relevant parts of the source sentence when generating the target translation. This contextual awareness enables NMT models to produce more accurate and contextually appropriate translations, capturing subtle nuances and preserving the meaning more effectively.

Scalability and parallelization

The transformer architecture, a key component of NMT models, allows for efficient parallelization during training and inference. This scalability has facilitated the training of larger models with more parameters, enabling NMT models to benefit from increased model capacity and achieve state-of-the-art results. The parallelization capability of transformers has significantly accelerated the training process and improved translation efficiency.

Ongoing research and innovation

Continuous research in NMT further enhances its capabilities through domain adaptation and transfer learning, ensuring that translation quality continues to improve even in specialized fields such as medicine.

4. MedEV dataset

The MedEV dataset is a high-quality Vietnamese - English parallel corpus specifically developed for the medical domain. It contains approximately 360,000 sentence pairs collected from diverse sources, split into 3 categories training (340,897 pairs), validation (8,982 pairs) and test (9006 pairs). Each sentence pair was carefully aligned and validated to ensure translation accuracy and domain relevance. MedEV provides a valuable benchmark for training and evaluating machine translation models, particularly for specialized Vietnamese-English medical translation tasks.

In the following section, I will evaluate the translation performance of two Neural Machine Translation (NMT) models mentioned in the MedEV paper. The BLEU score of each model will be calculated using the test split of the dataset, and the results will be compared with those reported in the original study.

5. Evaluation of envit5 translation model and VinAi Translate model

We evaluated two NMT models, envit5 translation and VinAi Translate, both of which are pre-trained for Vietnamese–English translation. The table below presents the BLEU scores of each model, obtained by testing on the test split of the MedEV dataset. At this stage, the models have not yet been fine-tuned on the MedEV dataset. The evaluation examines translation performance with respect to sentence length.

Due to computational constraints, it was not possible to test on the entire dataset; instead, the results were generated using a random sample of 100 sentences from the test set.

Model	Envit5-translation (w/o FT)					
	Vietnamese-to- English					
	<10	[10,20]	[20,30]	[30,40]	[40,50]	>50
Evaluation result on the paper with the full test set	38.72	41.77	42.75	43.73	44.08	42.59
Evaluation result on 100 random samples from the test set	**9.93**	**25.00**	**25.35**	**24.36**	**24.13**	**37.22**

Model	Envit5-translation (w/o FT)					
	English-to- Vietnamese					
	<10	[10,20]	[20,30]	[30,40]	[40,50]	>50
Evaluation result on the paper with the full test set	27.07	28.31	31.76	33.89	35.53	37.12
Evaluation result on 100 random samples from the test set	**29.33**	**40.77**	**44.08**	**45.48**	**42.32**	**43.53**

Model	Vinai-translate (w/o FT)					
	Vietnamese- to - English					
	<10	[10,20]	[20,30]	[30,40]	[40,50]	>50
Evaluation result on the paper with the full test set	28.81	30.99	33.03	34.36	36.01	38.01
Evaluation result on 100 random samples from the test set	**37.91**	**34.25**	**37.01**	**36.62**	**38.40**	**40.17**

Model	Vinai-translate (w/o FT)					
	English-to- Vietnamese					
	<10	[10,20]	[20,30]	[30,40]	[40,50]	>50
Evaluation result on the paper with the full test set	31.53	43.07	44.51	44.77	43.70	43.92
Evaluation result on 100 random samples from the test set	**44.45**	**47.80**	**48.52**	**50.81**	**53.94**	**58.11**

6. Observation from the evaluation result

From the evaluation results, we can conclude that, before fine-tuning on the MedEV dataset, both models envit5 and VinAi can perform English–Vietnamese and Vietnamese–English translations, but their performance in the medical domain remains limited.

For the **envit5 model**, Vietnamese-to-English translation achieved relatively low BLEU scores in our evaluation compared to the original study. For instance, for input sentences shorter than 10 words, the original study reported a BLEU score of **38.72**, while our evaluation yielded only **9.93**. This discrepancy is likely due to the small evaluation sample of 100.

For the **VinAI model**, our evaluation results were closer to those reported in the original study, although not identical. Notably, in English-to-Vietnamese translation, the BLEU score from our evaluation was even slightly higher than the original study, suggesting variability due to the limited sample size and testing conditions.

7. Future work

In the next report, we plan to extend our evaluation of the same two models using additional metrics, TER and METEOR, to provide a more comprehensive assessment of translation quality.

Furthermore, we will be looking for a methodology that allows us to evaluate the models on a larger set of samples, thereby obtaining results that more closely reflect those reported in the original study. We recognize that, with our current computing setup, we are limited to testing on a small number of samples and cannot fine-tune the models on the full dataset. Therefore, it is essential to develop a strategy to enhance our computational capacity, allowing both more extensive evaluation and thorough fine-tuning of the models.