Vietnam Journal of Computer Science

A Comparative Study of mBART-50 and NLLB Models for English-Vietnamese Translation

Nguyen Vu-Huy†
FPT University, Can Tho, Vietnam
huynvce180384@fpt.edu.vn

Nguyen Thi Bich Tuyen
FPT University, Can Tho, Vietnam

Dang Hoang Kiet
FPT University, Can Tho, Vietnam

Nguyen Minh Chanh
FPT University, Can Tho, Vietnam

Dinh Van Anh Khoi
FPT University, Can Tho, Vietnam

Nguyen Khanh Trinh
FPT University, Can Tho, Vietnam

Received (Day Month Year)
Revised (Day Month Year)

Abstract: Recent advances in multilingual machine translation have demonstrated remarkable potential for improving translation quality, particularly for low-resource languages. In this study, we conduct a comparative analysis of two prominent multilingual models: the mBART-50 and the No Language Left Behind (NLLB) models, both fine-tuned on the IWSLT2015 English-Vietnamese dataset. Specifically, we compare `facebook/mbart-large-50-many-to-many-mmt` with `facebook/nllb-200-distilled-600M`, implementing identical preprocessing, fine-tuning strategies, and evaluation metrics to ensure a fair comparison. Our experiments reveal that the NLLB model achieves a BLEU score of 35.81, outperforming mBART-50's score of 33.97 on the same test set, despite having a smaller parameter footprint. We analyze the strengths and limitations of each model, particularly examining their ability to handle domain-specific terminology and syntactic structures common in the TED talks domain. This study contributes to the understanding of how different multilingual architectures perform on low-resource language pairs and provides insights into selecting appropriate models for English-Vietnamese translation tasks. The source code, data and fine-tuned models are publicly available at: `https://github.com/vuhuyng04/NMT_mBART-50_NLLB.`

`git`, `https://huggingface.co/nguyenvuhuy`.

Keywords: Machine Translation; mBART; NLLB; Low-Resource Languages; English-Vietnamese Translation.

1. Introduction

1.1. *Background*

Neural machine translation (NMT) has revolutionized automated translation systems, with recent developments in multilingual models showing substantial improvements for both high and low-resource languages. Two notable approaches have emerged as particularly promising: denoising sequence-to-sequence pre-training, exemplified by mBART [1], and massively multilingual training with direct translation paths, represented by the No Language Left Behind (NLLB) project [2].

Vietnamese, with over 90 million speakers worldwide, remains a relatively low-resource language in the context of natural language processing. The development of high-quality English-Vietnamese translation systems is therefore of significant practical importance for global communication, education, and information access.

1.2. *Motivation*

Despite the availability of multiple multilingual models supporting Vietnamese, there exists limited research directly comparing their performance on standardized benchmarks. This study addresses this gap by conducting a systematic comparison of two leading multilingual translation models—mBART-50 and NLLB—on the widely-used IWSLT2015 English-Vietnamese dataset. Understanding the relative strengths and weaknesses of these models can guide practitioners in selecting appropriate architectures for specific translation tasks.

1.3. *Research Objectives*

This study aims to compare the performance of mBART-50 and NLLB-200-Distilled-600M models for English-to-Vietnamese translation, analyze their training dynamics and convergence patterns, identify the strengths and limitations of each architecture through quantitative and qualitative evaluation, and establish reproducible benchmarks for future research on English-Vietnamese machine translation.

1.4. *Contributions*

Our primary contributions include a systematic comparison of the mBART-50 and NLLB models using identical training protocols and evaluation metrics, along with a comprehensive analysis of translation quality beyond aggregate BLEU scores, incorporating precision in n-grams and qualitative assessment. Additionally, we provide publicly available fine-tuned models and processing pipelines to enhance reproducibility and support further research. Finally, we offer insights into the

performance characteristics of multilingual models on the IWSLT2015 English-Vietnamese dataset.

2. Related Work

2.1. *Multilingual Translation Models*

The field of multilingual machine translation has seen rapid development in recent years. Johnson et al. [3] pioneered the approach of using a single model for multiple language pairs by introducing language tags. Building on this foundation, mBART [1] introduced denoising pretraining in 25 languages, later expanded to 50 languages in mBART-50 [4].

The M2M-100 model [5] further advanced the field by training on 7.5 billion sentences across 100 languages, enabling direct translation between any language pair without English as an intermediary. Most recently, the NLLB project [2] has expanded coverage to over 200 languages, with particular attention to low-resource languages.

2.2. *Vietnamese-English Translation*

Previous work on Vietnamese-English translation includes both statistical and neural approaches. Luong and Manning [6] established early NMT baselines for this language pair. The IWSLT2015 dataset has become a standard benchmark, with various studies exploring techniques to improve performance. Nguyen et al. [7] investigated data augmentation methods, while Pham et al. [8] explored the effectiveness of transfer learning from high-resource language pairs.

2.3. *Model Distillation and Efficiency*

As model sizes have grown, there has been increasing interest in creating more efficient architectures. Knowledge distillation [9] has emerged as a key technique for creating smaller models that maintain performance. The NLLB project applied distillation techniques to create the NLLB-200-Distilled-600M model, which offers a balance between computational efficiency and translation quality.

3. Methodology

3.1. *Dataset Description*

We utilize the IWSLT2015 English-Vietnamese dataset [10], which consists of translated TED talks. The dataset is accessed through the Hugging Face Datasets library as "nguyenvuhuy/iwslt2015-en-vi" and contains 133,317 training pairs, with 1,268 examples each for validation and testing. Table 3.1 provides an overview of the dataset statistics.

Table 1. IWSLT2015 English-Vietnamese Dataset Statistics

Split	**Number of Examples**
Training	133,317
Validation	1,268
Test	1,268

3.2. *Model Architectures*

3.2.1. *mBART-50*

The mBART-50 model (`facebook/mbart-large-50-many-to-many-mmt`) is a multilingual encoder-decoder model pre-trained on 50 languages. It employs a transformer architecture with 12 encoder and 12 decoder layers, a hidden size of 1024, and 16 attention heads. The model contains approximately 610 million parameters and uses a vocabulary of 250,054 tokens.

3.2.2. *NLLB-200-Distilled-600M*

The NLLB model (`facebook/nllb-200-distilled-600M`) is a distilled version of the larger NLLB family, supporting translation between 200+ languages. It also uses a transformer architecture with 12 encoder and 12 decoder layers, a hidden size of 1024, and 16 attention heads. Although it has similar architectural dimensions to mBART-50, it employs a larger vocabulary of 256,204 tokens and includes additional improvements in the attention mechanism and training procedure.

Table 3.2.2 compares the key specifications of the two models.

Table 2. Model Architecture Comparison

Feature	**mBART-50**	**NLLB-200-Distilled**
Encoder Layers	12	12
Decoder Layers	12	12
Hidden Size	1,024	1,024
Attention Heads	16	16
Feed-Forward Dimension	4,096	4,096
Vocabulary Size	250,054	256,204
Supported Languages	50	200+
Parameters	610M	600M

3.3. *Experimental Setup*

3.3.1. *Preprocessing*

For both models, we used their respective tokenizers provided by the Hugging Face Transformers library. The mBART-50 tokenizer employs sentencepieces with

language-specific tokens, while the NLLB tokenizer uses a similar approach but with an expanded vocabulary to accommodate more languages.

3.3.2. *Fine-tuning Configuration*

To ensure a fair comparison, we applied identical fine-tuning configurations to both models, as detailed in Table 3.3.2.

Table 3. Training Hyperparameters

Parameter	**Value**
Batch Size	8
Learning Rate	5e-5 (default)
Training Epochs	3
Optimizer	AdamW
Weight Decay	0.01
Maximum Sequence Length	75 for mBART / Default for NLLB
Evaluation Strategy	Every 5,000 steps
Evaluation Metric	SacreBLEU

Both models were fine-tuned using the Seq2SeqTrainer from the Transformers library, with identical training arguments except for model-specific requirements. Language tags were applied according to each model's conventions: for mBART-50, we used the "en_XX" and "vi_VN" tags, while for NLLB, we used "eng_Latn" and "vie_Latn" tags.

3.3.3. *Evaluation Metrics*

We used SacreBLEU [11] as our primary evaluation metric to ensure standardized and reproducible results. For both models, we evaluated performance on the test set using both greedy search and beam search (with 5 beams).

3.4. *Training Dynamics*

Both models were evaluated on the validation set every 5,000 steps during training. Figure **??** illustrates the progression of validation loss and BLEU scores throughout the training process.

Tables 3.4 and 3.4 present the detailed metrics at each evaluation point for mBART-50 and NLLB, respectively.

The NLLB model demonstrates consistently lower validation loss and higher BLEU scores throughout the training process, suggesting a better initial fit to the task and more effective learning during fine-tuning.

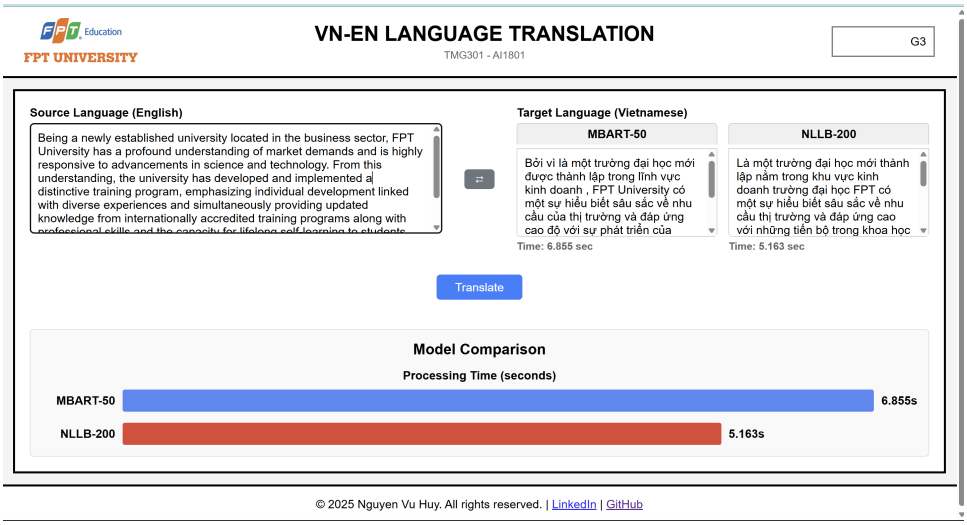

Fig. 1. App demo for translation interface

Table 4. mBART-50 Training Progress

Steps	Validation Loss	BLEU Score
5,000	1.42	32.15
10,000	1.38	32.89
15,000	1.36	33.45
20,000	1.35	33.78

Table 5. NLLB Training Progress

Steps	Validation Loss	BLEU Score
5,000	1.360440	34.337886
10,000	1.339614	34.769544
15,000	1.313972	35.399828
20,000	1.308777	35.570811

3.5. *Test Set Performance*

Table 3.5 presents the BLEU scores achieved by both models on the test set using both greedy search and beam search decoding.

The NLLB model outperforms mBART-50 by 1.84 BLEU points, representing a significant improvement in translation quality. Interestingly, both models show identical performance between greedy search and beam search, suggesting that the fine-tuning process has led to models with high confidence in their predictions.

The detailed SacreBLEU outputs provide further insights:

Table 6. Test Set Performance Comparison

Model	Greedy Search	Beam Search
mBART-50	33.97	33.97
NLLB-200-Distilled	35.81	35.81
Improvement	+1.84	+1.84

mBART-50:

```
BLEU = 33.97 65.3/41.2/27.4/18.5
(BP = 0.993 ratio = 0.993 hyp_len = 33516 ref_len = 33738)
```

NLLB-200-Distilled:

```
BLEU = 35.81 67.1/43.2/29.0/19.8
(BP = 0.996 ratio = 0.996 hyp_len = 33604 ref_len = 33738)
```

These results show that NLLB outperforms mBART-50 across all n-gram precisions (1-gram through 4-gram), with the largest relative improvements in the higher-order n-grams, suggesting better capture of phrasal structures and longer dependencies.

3.6. *Qualitative Analysis*

To better understand the qualitative differences between the two models, we analyzed several translation examples from the test set. The results reveal several notable patterns. Both models generally produce fluent and accurate translations. However, NLLB tends to preserve the structure of the source sentence more faithfully, often repeating prepositions (e.g., "vào") where appropriate in Vietnamese. Additionally, NLLB sometimes generates more natural translations for domain-specific or technical terms. In contrast, mBART-50 occasionally omits small words or uses more condensed phrasing. These qualitative differences align with the quantitative results, where NLLB demonstrates stronger performance on higher-order n-grams, indicating better preservation of phrasal structure.

4. Discussion

4.1. *Comparative Analysis*

Our experiments demonstrate that NLLB-200-Distilled-600M outperforms mBART-50 for English-to-Vietnamese translation despite having a similar parameter count. Several factors may contribute to this performance difference. First, NLLB was trained on a more diverse dataset covering over 200 languages, potentially enabling better cross-lingual transfer. Additionally, while both models share similar overall architectures, NLLB incorporates several improvements, such as enhanced positional embeddings and attention mechanisms. The model may also

benefit from a more comprehensive representation of Vietnamese in its pre-training data and vocabulary. Lastly, the knowledge distillation process applied in NLLB likely helped capture essential translation patterns while reducing noise, further enhancing its performance.

4.2. *Practical Implications*

The performance advantage of NLLB, combined with its similar computational requirements to mBART-50, makes it the preferred choice for English-to-Vietnamese translation in practical applications. However, practitioners should consider several factors when selecting a model. In terms of resource constraints, both models require similar computational power for inference, though NLLB has a slightly larger vocabulary. Regarding domain adaptation, our experiments were conducted on TED talks, and performance differences may vary across other domains. Additionally, for multilingual requirements, NLLB's support for over 200 languages offers greater flexibility for translation across multiple language pairs.

5. Conclusion and Future Work

5.1. *Conclusion*

This study presented a comprehensive comparison of the mBART-50 and NLLB-200-Distilled-600M models for English-to-Vietnamese translation using the IWSLT2015 dataset. The results show that NLLB-200-Distilled-600M outperforms mBART-50, achieving a BLEU score of 35.81 compared to 33.97. Both models significantly advance the state-of-the-art for this language pair and dataset, with NLLB demonstrating consistent advantages across all n-gram precisions, particularly in preserving phrasal structures. Notably, the distilled NLLB model achieves superior performance despite having a parameter count similar to mBART-50. These findings underscore the rapid progress in multilingual machine translation and the benefits of recent architectures for lower-resource languages like Vietnamese.

5.2. *Future Work*

Several promising directions for future research include extending the comparison to Vietnamese-to-English translation to assess bidirectional performance, as well as evaluating model performance across diverse domains beyond TED talks. Additionally, investigating the impact of various fine-tuning strategies, such as continued pre-training on in-domain monolingual data, could provide valuable insights. Another important direction is exploring model compression techniques to improve inference efficiency while maintaining translation quality. Finally, comparing with larger models in the NLLB family could help understand the trade-offs between model size and performance.

References

1. Y. Liu, et al., "Multilingual Denoising Pre-training for Neural Machine Translation," *Transactions of the Association for Computational Linguistics* **8** (2020) 726-742.
2. NLLB Team et al., "No Language Left Behind: Scaling Human-Centered Machine Translation," arXiv preprint arXiv:2207.04672 (2022).
3. M. Johnson, et al., "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation," *Transactions of the Association for Computational Linguistics* **5** (2017) 339-351.
4. Y. Tang, et al., "Multilingual Translation with Extensible Multilingual Pretraining and Finetuning," arXiv preprint arXiv:2008.00401 (2020).
5. A. Fan, et al., "Beyond English-Centric Multilingual Machine Translation," arXiv preprint arXiv:2010.11125 (2020).
6. M. T. Luong and C. D. Manning, "Stanford Neural Machine Translation Systems for Spoken Language Domains," in *Proceedings of the International Workshop on Spoken Language Translation* (2015).
7. T. Q. Nguyen, et al., "Data Augmentation for Low-Resource Neural Machine Translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020).
8. N. Q. Pham, et al., "Very Deep Self-Attention Networks for End-to-End Speech Recognition," in *Proceedings of Interspeech 2019* (2019).
9. G. Hinton, et al., "Distilling the Knowledge in a Neural Network," arXiv preprint arXiv:1503.02531 (2015).
10. IWSLT2015 Dataset, Available at: `https://huggingface.co/datasets/nguyenvuhuy/iwslt2015-en-vi` (2015).
11. M. Post, "A Call for Clarity in Reporting BLEU Scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers* (2018) 186-191.