International IT Journal of Research (IITJR), ISSN: 3007-6706
Volume 4, Issue 1, January-March 2026
Available online at: https://itjournal.org/index.php

# Low-Resource Neural Machine Translation for Southeast Asian Languages Using Cross-Lingual Transfer and Subword Regularization

**Dr. Thao Phuong Le**

Assistant Professor, School of Information and Communication Technology, Hanoi University of Science and Technology, Vietnam

## ABSTRACT

**Neural Machine Translation (NMT) has achieved remarkable progress in facilitating multilingual communication; however, its effectiveness remains limited for low-resource languages due to insufficient parallel corpora, limited linguistic resources, and significant variations in vocabulary, morphology, and grammatical structures. Many Southeast Asian languages, including Khmer, Lao, Burmese, Tagalog, Vietnamese, and several regional languages, continue to face challenges in developing high-quality machine translation systems because of the scarcity of annotated bilingual datasets. This study explores an enhanced low-resource Neural Machine Translation framework that integrates cross-lingual transfer learning and subword regularization to improve translation accuracy, linguistic representation, and generalization capabilities. The proposed approach leverages multilingual pretrained language models and high-resource language pairs to transfer contextual and syntactic knowledge to low-resource target languages.**

**Additionally, subword regularization is employed to generate diverse segmentation patterns during training, reducing vocabulary sparsity and improving the handling of rare words, compound expressions, and morphologically complex linguistic structures. The framework incorporates multilingual encoder-decoder architectures, shared subword vocabularies, transfer-based parameter initialization, and data augmentation techniques to maximize the utilization of limited parallel training resources. Its performance can be evaluated using established translation metrics, including BLEU, chrF, and COMET, alongside computational efficiency and cross-domain generalization measures. The proposed methodology aims to demonstrate the potential of combining cross-lingual knowledge transfer with subword-level variation to address data scarcity and improve translation robustness across linguistically diverse Southeast Asian languages. This research contributes toward developing scalable, inclusive, and resource-efficient multilingual translation technologies, supporting digital accessibility, preservation of regional languages, and broader participation in global information exchange.**

**Keywords: Neural Machine Translation (NMT), Low-Resource Languages, Cross-Lingual Transfer Learning, Subword Regularization, Southeast Asian Languages**

## INTRODUCTION

The rapid development of Artificial Intelligence (AI), Natural Language Processing (NLP), and deep learning technologies has significantly transformed the field of machine translation. Neural Machine Translation (NMT) has emerged as one of the most effective approaches for automatically translating text between different languages by utilizing deep neural networks to learn complex linguistic patterns, semantic relationships, and contextual representations. Unlike traditional rule-based and statistical machine translation systems, NMT models employ encoder-decoder architectures, attention mechanisms, and Transformer-based networks to generate more fluent and contextually appropriate translations. Despite these advancements, the performance of neural machine translation systems remains highly dependent on the availability of large-scale parallel corpora and high-quality linguistic resources.

One of the major challenges in contemporary machine translation research is the limited availability of parallel training data for low-resource languages. While languages such as English, Chinese, French, and Spanish benefit from extensive digital resources, numerous Southeast Asian languages continue to receive comparatively limited computational and linguistic support. Languages such as Khmer, Lao, Burmese, Tagalog, Javanese, Sundanese, and other regional languages face difficulties associated with insufficient bilingual datasets, complex morphological structures, diverse writing systems, and limited standardized linguistic resources. These limitations negatively affect translation accuracy, vocabulary coverage, grammatical consistency, and the ability of machine translation systems to understand context-dependent expressions.

## PROPOSED MODELS AND METHODOLOGIES

### 2.1 Proposed Research Framework

The proposed research introduces an enhanced Low-Resource Neural Machine Translation (LR-NMT) framework for Southeast Asian languages by integrating cross-lingual transfer learning, multilingual Transformer architectures, and subword regularization. The primary objective is to improve translation accuracy, lexical representation, and generalization when only limited parallel bilingual datasets are available.

The framework utilizes linguistic knowledge transferred from high-resource languages to low-resource language pairs. It also applies stochastic subword segmentation to improve the representation of rare words, compound expressions, and morphologically complex terms. The overall architecture consists of data collection, preprocessing, multilingual representation learning, cross-lingual knowledge transfer, subword regularization, model training, and translation evaluation.

### 2.2 Dataset Collection and Preprocessing

The first stage involves collecting parallel bilingual datasets for selected Southeast Asian languages. The datasets may include English–Khmer, English–Lao, English–Burmese, English–Tagalog, and English–Vietnamese language pairs.
Potential data sources include OPUS, OPUS-100, FLORES-200, and other publicly available multilingual language resources. High-resource language datasets are also incorporated to facilitate cross-lingual knowledge transfer.
The preprocessing stage includes:

- **Text normalization:** Standardizing Unicode characters, punctuation, and language-specific writing conventions.
- **Sentence alignment:** Ensuring accurate correspondence between source and target sentences.
- **Noise removal:** Eliminating duplicate, corrupted, and incorrectly aligned samples.
- **Language identification:** Verifying the language identity of individual text segments.
- **Tokenization:** Converting textual information into suitable subword representations.
- **Dataset balancing:** Reducing excessive dominance of high-resource language pairs during multilingual training.

The dataset is divided into training, validation, and testing subsets to ensure reliable model development and evaluation.

### 2.3 Multilingual Transformer-Based Translation Model

The proposed framework employs a multilingual Transformer encoder-decoder architecture as its primary translation model. Unlike conventional recurrent neural networks, Transformer architectures utilize self-attention mechanisms to capture long-range contextual dependencies and relationships between words.
The encoder converts the source-language sentence into contextual representations, while the decoder generates the corresponding target-language translation through autoregressive prediction.
The attention mechanism is represented as:
\[ Attention(Q,K,V)= Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \]
Where:

- \(Q\) represents query vectors.
- \(K\) represents key vectors.
- \(V\) represents value vectors.
- \(d_k\) denotes the dimensionality of key vectors.

The multilingual architecture incorporates shared embedding spaces to capture similarities across different languages. Language-specific identifiers are introduced to guide the translation model toward the appropriate target language.
This architecture supports knowledge sharing across multiple language pairs while maintaining contextual understanding and grammatical consistency.

### 2.4 Cross-Lingual Transfer Learning Mechanism

Cross-lingual transfer learning represents one of the central components of the proposed methodology. It enables low-resource languages to benefit from linguistic representations learned through high-resource language datasets.
The proposed mechanism initially trains or initializes a multilingual translation model using high-resource language pairs. Subsequently, the learned parameters are transferred to low-resource language translation tasks through fine-tuning.
The process consists of the following stages:

1. **High-resource pretraining:** Training the multilingual model using large-scale parallel corpora.
2. **Shared representation learning:** Developing common semantic and contextual representations across languages.
3. **Parameter transfer:** Reusing pretrained encoder and decoder parameters for low-resource language pairs.
4. **Language-specific adaptation:** Fine-tuning the model using available low-resource bilingual datasets.

5. **Performance optimization:** Adjusting learning rates and training parameters to minimize translation errors.

### 2.5 Subword Regularization Technique

Subword regularization is incorporated to improve vocabulary representation and address the challenges associated with rare words and limited training data.

Conventional tokenization approaches generally generate fixed segmentation patterns. However, the proposed methodology introduces stochastic segmentation, allowing the same sentence to be represented through multiple possible subword combinations during training.

The Unigram Language Model and SentencePiece tokenization techniques can be employed to generate alternative subword segmentations.

The probability of selecting a segmentation \(s\) for a sentence \(x\) can be represented as:

\[ P(s|x)= \frac{P(s)^\alpha} {\sum_{s'\in S(x)}P(s')^\alpha} \]

Where:

- \(s\) represents a possible subword segmentation.
- \(S(x)\) represents the set of valid segmentations.
- \(P(s)\) represents the probability assigned to a segmentation.
- \(\alpha\) controls the degree of segmentation randomness.

By exposing the model to different segmentation patterns, subword regularization improves lexical flexibility and reduces overdependence on fixed vocabulary representations.

## RESULTS & ANALYSIS

### 3.1 Overview of Experimental Results

The proposed Low-Resource Neural Machine Translation (LR-NMT) framework integrates cross-lingual transfer learning and subword regularization to improve translation performance across Southeast Asian languages. The experimental evaluation is designed to investigate the effectiveness of the proposed approach under different data availability conditions.

The study considers selected language pairs, including English–Khmer, English–Lao, English–Burmese, English–Tagalog, and English–Vietnamese. The proposed framework is compared with conventional Neural Machine Translation, multilingual Transformer models, cross-lingual transfer-based models, and subword-regularized translation systems.

The evaluation focuses on translation accuracy, semantic preservation, vocabulary coverage, robustness, and computational efficiency. BLEU, chrF, and COMET metrics are considered to assess translation quality.

***Note:*** *The following numerical results are illustrative hypothetical results for research-paper presentation. They are not actual experimental measurements and should be replaced with results obtained from model implementation and testing.*

### 3.2 Comparative Performance of Different Models

The following table presents an illustrative comparison of five translation architectures.

| Model | BLEU Score | chrF Score | COMET Score |
|---|---|---|---|
| Conventional NMT | 18.6 | 42.3 | 0.48 |
| Multilingual Transformer | 22.4 | 46.8 | 0.55 |
| Cross-Lingual Transfer NMT | 26.7 | 51.2 | 0.62 |
| Subword-Regularized NMT | 24.9 | 49.6 | 0.59 |
| **Proposed Combined Model** | **30.8** | **55.7** | **0.69** |

The illustrative results indicate that the proposed combined architecture achieves the highest performance across all three evaluation metrics. The integration of cross-lingual transfer and subword regularization demonstrates the potential to improve translation quality beyond the individual techniques.

The conventional NMT model records comparatively lower performance because of its dependence on limited parallel training data. In contrast, cross-lingual transfer improves the availability of contextual linguistic representations, while subword regularization provides greater flexibility in handling rare vocabulary.

### 3.3 Translation Performance across Southeast Asian Languages

The effectiveness of the proposed framework is further examined across different Southeast Asian language pairs.

| Language Pair | Baseline BLEU | Proposed BLEU | Improvement |
|---|---|---|---|
| English–Khmer | 15.8 | 27.6 | 11.8 |
| English–Lao | 16.4 | 28.3 | 11.9 |
| English–Burmese | 17.2 | 29.1 | 11.9 |
| English–Tagalog | 23.5 | 34.2 | 10.7 |
| English–Vietnamese | 25.1 | 35.6 | 10.5 |

The results suggest that the proposed approach may provide substantial benefits for languages with limited bilingual training resources. Khmer and Lao demonstrate relatively large improvements over their baseline configurations. This highlights the potential value of transferring multilingual representations to language pairs with limited parallel data.

Tagalog and Vietnamese achieve comparatively higher absolute BLEU scores, potentially reflecting differences in available training resources, linguistic characteristics, and dataset quality.

### 3.4 Impact of Cross-Lingual Transfer Learning

Cross-lingual transfer learning is evaluated to understand its contribution to low-resource translation performance.

| Training Data Availability | Baseline BLEU | Cross-Lingual BLEU |
|---|---|---|
| 10% Dataset | 10.8 | 19.7 |
| 25% Dataset | 14.2 | 23.5 |
| 50% Dataset | 17.9 | 26.8 |
| 75% Dataset | 20.6 | 29.1 |
| 100% Dataset | 23.4 | 31.2 |

The comparison indicates that cross-lingual transfer learning can provide considerable advantages when only a small proportion of parallel training data is available.

At the 10% data level, the transferred model achieves a BLEU score of 19.7 compared with 10.8 for conventional NMT. This suggests that multilingual pretraining can compensate partially for the absence of extensive bilingual training examples.

As dataset availability increases, both approaches improve. However, the cross-lingual model maintains a performance advantage across all experimental configurations.

### 3.5 Impact of Subword Regularization

Subword regularization is investigated to determine its effectiveness in improving lexical representation and handling rare words.

| Tokenization Method | BLEU | chrF | Rare-Word Accuracy (%) |
|---|---|---|---|
| Word-Level Tokenization | 17.3 | 40.2 | 51.4 |
| BPE Tokenization | 22.4 | 46.8 | 64.7 |
| SentencePiece Unigram | 24.1 | 48.9 | 68.5 |
| Subword Regularization | 26.2 | 51.3 | 74.8 |

The results indicate that stochastic subword segmentation has the potential to improve translation quality compared with fixed word-level and subword tokenization methods.

## COMPARATIVE ANALYSIS IN TABULAR FORM

The proposed framework combines cross-lingual transfer learning, multilingual Transformer architecture, and subword regularization to improve translation performance for low-resource Southeast Asian languages.

### 4.1 Comparative Analysis of Different Translation Models

| S. No. | Translation Model | BLEU Score | chrF Score | COMET Score |
|---|---|---|---|---|
| 1 | Conventional NMT | 18.6 | 42.3 | 0.48 |
| 2 | Multilingual Transformer | 22.4 | 46.8 | 0.55 |
| 3 | Cross-Lingual Transfer NMT | 26.7 | 51.2 | 0.62 |
| 4 | Subword-Regularized NMT | 24.9 | 49.6 | 0.59 |
| 5 | **Proposed Combined Model** | **30.8** | **55.7** | **0.69** |

**Analysis:** The proposed combined model achieves the highest scores across all three evaluation metrics. This indicates the potential benefits of integrating cross-lingual knowledge transfer with stochastic subword segmentation.

### 4.2 Comparative Analysis of Technical Features

| Parameters | Conventional NMT | Multilingual NMT | Cross-Lingual NMT | Proposed Model |
|---|---|---|---|---|
| Training Data Requirement | High | High | Medium | Low–Medium |
| Knowledge Transfer | No | Partial | Yes | Advanced |
| Subword Regularization | Limited | Optional | Optional | Integrated |
| Rare-Word Handling | Limited | Moderate | Good | Strong |
| Multilingual Support | Limited | High | High | High |
| Vocabulary Coverage | Moderate | Good | High | Improved |
| Generalization | Limited | Moderate | Good | Improved |
| Scalability | Moderate | High | High | High |

### 4.3 Comparative Analysis across Southeast Asian Languages

| Language Pair | Baseline BLEU | Proposed BLEU | Improvement |
|---|---|---|---|
| English–Khmer | 15.8 | 27.6 | +11.8 |
| English–Lao | 16.4 | 28.3 | +11.9 |
| English–Burmese | 17.2 | 29.1 | +11.9 |
| English–Tagalog | 23.5 | 34.2 | +10.7 |
| English–Vietnamese | 25.1 | 35.6 | +10.5 |

**Analysis:** The proposed approach demonstrates improvements across all five illustrative language-pair experiments. The results suggest that cross-lingual transfer can be particularly beneficial for language pairs with limited parallel training resources.

## LIMITATIONS & DRAWBACKS

Although the proposed Low-Resource Neural Machine Translation (LR-NMT) framework combines cross-lingual transfer learning and subword regularization to improve translation performance for Southeast Asian languages, several limitations and challenges may affect its practical implementation. The integration of multilingual Transformer architectures, transferred linguistic representations, and stochastic tokenization offers considerable potential; however, its effectiveness depends on dataset quality, language similarity, computational resources, and linguistic diversity.

The major limitations and drawbacks are discussed below.

### 5.1 Limited Availability of Parallel Datasets

One of the primary challenges in low-resource Neural Machine Translation is the scarcity of high-quality parallel bilingual datasets. Many Southeast Asian languages have insufficient translated documents, standardized corpora, and publicly accessible linguistic resources.

The lack of parallel data can negatively affect model training, vocabulary learning, and translation accuracy. Even with cross-lingual transfer learning, the model may struggle to understand language-specific expressions and complex grammatical structures.

### 5.2 Linguistic Differences between Languages

Cross-lingual transfer learning depends partly on the ability to share useful linguistic representations across different languages. However, Southeast Asian languages exhibit considerable variations in syntax, morphology, sentence structure, and writing systems.

For example, Khmer, Burmese, Vietnamese, Lao, and Tagalog have distinct linguistic characteristics. Knowledge transferred from a high-resource language may not always be directly applicable to another language.

Consequently, negative transfer may occur, reducing translation accuracy instead of improving it.

### 5.3 Subword Segmentation Complexity

Subword regularization improves vocabulary flexibility by introducing alternative segmentation patterns during training. However, excessive segmentation variation can sometimes produce inconsistent lexical representations.

The selection of segmentation probabilities and sampling parameters significantly influences model performance. Inappropriate configurations may divide meaningful linguistic units into less informative subword fragments.

This limitation is particularly relevant for languages with complex morphological structures and writing conventions.

## CONCLUSION

The present research, titled **"Low-Resource Neural Machine Translation for Southeast Asian Languages Using Cross-Lingual Transfer and Subword Regularization,"** investigates an advanced approach to addressing the challenges associated with machine translation for linguistically diverse and underrepresented languages. Although Neural Machine Translation (NMT) has achieved significant advancements through deep learning and Transformer-based architectures, its effectiveness remains constrained by the limited availability of parallel bilingual datasets, complex linguistic structures, vocabulary sparsity, and insufficient computational resources. These challenges are particularly significant for Southeast Asian languages, including Khmer, Lao, Burmese, Tagalog, and Vietnamese.

This research proposes an enhanced low-resource Neural Machine Translation framework that integrates multilingual Transformer architectures, cross-lingual transfer learning, and subword regularization. Cross-lingual transfer learning enables the utilization of linguistic knowledge obtained from high-resource languages, thereby reducing dependence on extensive parallel training datasets. Meanwhile, subword regularization introduces alternative segmentation patterns during training, improving vocabulary representation and the handling of rare words, compound expressions, and morphologically complex linguistic structures.

The proposed methodology also incorporates multilingual representation learning, data preprocessing, back-translation, synthetic data generation, and optimized model training to improve translation robustness and generalization. The combined approach provides a systematic framework for developing translation systems capable of learning from limited bilingual resources while maintaining contextual and semantic relationships across different languages.

The illustrative comparative analysis indicates that the proposed combined model can achieve improved translation performance compared with conventional NMT, multilingual Transformer, cross-lingual transfer, and subword-regularized approaches considered independently. In the hypothetical experimental evaluation, the combined model achieves a BLEU score of 30.8, a chrF score of 55.7, and a COMET score of 0.69. These values demonstrate the intended evaluation framework and require verification through actual experimental implementation before being interpreted as empirical findings.

## REFERENCES

1. Aharoni, R., Johnson, M., & Firat, O. (2019). Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3874–3884). Association for Computational Linguistics.
2. Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G., Cherry, C., Macherey, W., Chen, Z., & Wu, Y. (2019). Massively multilingual neural machine translation in the wild: Findings and challenges. arXiv. https://arxiv.org/abs/1907.05019
3. Beloucif, M., Gonzalez, A. V., Bollmann, M., & Søgaard, A. (2019). Naive regularizers for low-resource neural machine translation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*.
4. Fadaee, M., Bisazza, A., & Monz, C. (2017). Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers, pp. 567–573). Association for Computational Linguistics.
5. Goyal, N., Gao, C., Chaudhary, V., Chen, P. J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., & Fan, A. (2022). The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, *10*, 522–538.
6. Ha, T. L., Niehues, J., & Waibel, A. (2016). Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th International Conference on Spoken Language Translation*.

7. Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., & Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics, 5*, 339–351.
8. Karakanta, A., Dehdari, J., & van Genabith, J. (2018). Neural machine translation for low-resource languages without parallel corpora. *Machine Translation, 32*, 167–189.
9. Kim, Y., Gao, Y., & Ney, H. (2019). Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1246–1257). Association for Computational Linguistics.
10. Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.
11. Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers, pp. 66–75). Association for Computational Linguistics.
12. Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 66–71). Association for Computational Linguistics.
13. Meyer, F., & Buys, J. (2024). A systematic analysis of subwords and cross-lingual transfer in multilingual translation. In *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 2194–2200). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.141
14. Neubig, G., & Hu, J. (2018). Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 875–880). Association for Computational Linguistics.
15. NLLB Team. (2024). Scaling neural machine translation to 200 languages. *Nature, 630*, 841–846. https://doi.org/10.1038/s41586-024-07335-x
16. Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., & Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations* (pp. 48–53). Association for Computational Linguistics.
17. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics.
18. Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the 10th Workshop on Statistical Machine Translation* (pp. 392–395). Association for Computational Linguistics.
19. Ranathunga, S., Lee, E.-S. A., Skenduli, M. P., Shekhar, R., Alam, M., & Kaur, R. (2023). Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys, 55*(11), Article 229. https://doi.org/10.1145/3567592
20. Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 2685–2702). Association for Computational Linguistics.