JTE HCM-UTE
ISSN: 2615-9740

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Engineering
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

Multilingual Neural Machine Translation for Asian Language Treebank

Hong Buu Long Nguyen*, Thanh Tung Vu
University of Science, VNU-HCM, Vietnam

* Corresponding author. *Email: nhblong@fit.hcmus.edu.vn*

ARTICLE INFO

Received: 15/12/2025
Revised: 06/02/2026
Accepted: 25/02/2026
Published: 28/02/2026

KEYWORDS

Multilingual;
Neural Machine Translation;
Asian Languages;
Low Resources;
Asian Language Treebank.

ABSTRACT

This study examines multilingual neural machine translation (MNMT) for a diverse group of low-resource Asian languages-Bengali, Filipino, Indonesian, Japanese, Khmer, Malay, and Vietnamese-which differ substantially in linguistic families, writing systems, and typology. This paper evaluates state-of-the-art MNMT systems and introduces a Compact & Language-Sensitive MNMT model designed to improve translation performance while reducing computational cost. The proposed approach shares parameters through a compact multilingual representation, and enhances language discrimination using language-sensitive embeddings, a language-sensitive discriminator, and an adaptive cross-attention mechanism that selects attention parameters based on specific language pairs. Integrated with a multi-stage fine-tuning strategy, this model effectively strengthens cross-lingual transfer while maintaining robust language-specific representations. Experiments on the ALT multi-parallel corpus and the KFTT English-Japanese dataset demonstrate that multilingual models significantly outperform single-language NMT baselines. Despite its smaller size, the proposed Compact & Language-Sensitive MNMT achieves competitive or superior BLEU scores compared to Google's MNMT, confirming the effectiveness of guided parameter sharing and language-sensitive training. These results highlight the value of compact multilingual architectures and multi-parallel datasets for advancing low-resource Asian machine translation.

Doi: https://doi.org/10.54644/jte.2026.2047

1. Introduction

Neural Machine Translation (NMT) [1], [2] constitutes a modern paradigm in machine translation (MT), driven by recent advances in sequence-to-sequence learning frameworks [3], [4]. Over the past decade, NMT has achieved substantial performance gains and has attracted considerable attention from both academia and industry. In practical applications, the canonical NMT architecture is the attention-based encoder-decoder framework. Within this framework, the encoder maps a source sequence to a continuous vector representation, and the decoder generates the target sequence in an autoregressive manner using neural networks. Both the encoder and the decoder can be implemented using a variety of neural architectures, including Recurrent Neural Networks (RNNs) [4], [5], Convolutional Neural Networks (CNNs) [6], and self-attention-based models [7]. The attention mechanism, situated between the encoder and the decoder, allows the decoder to selectively attend to relevant parts of the encoded source sequence at each decoding step, thereby mitigating the limitations of fixed-length vector representations.

Early research in neural machine translation primarily focused on the development of bilingual translation systems; however, subsequent studies have demonstrated that NMT frameworks can effectively exploit multilingual data, leading to substantial growth in research on systems that support multiple language pairs [8]-[13]. These systems, commonly referred to as multilingual NMT (MNMT),

enable knowledge transfer across languages within a unified model. Such cross-lingual transfer has been shown to benefit not only low-resource languages, which suffer from limited parallel corpora or linguistic resources and can leverage information from related languages [14], but also high-resource languages with access to large-scale parallel datasets [12]. Owing to these advantages, multilingual training has increasingly been investigated as a form of data augmentation to enhance translation performance. Nevertheless, most existing studies predominantly focus on European languages, with comparatively limited attention given to languages from other regions. European languages are often closely related, belonging to the same or similar language families or sharing writing systems based on the Latin alphabet, which facilitates vocabulary and representation sharing within NMT models. Consequently, conclusions drawn from experiments on such closely related languages may not generalize to linguistically distant language pairs, such as many Asian languages.

This paper investigates and evaluates the effects of multilingual neural machine translation (MNMT) on several selected low-resource Asian language pairs, including Bengali (Bn), Filipino (Tl), Indonesian (Id), Japanese (Ja), Khmer (Km), Malay (Ms), and Vietnamese (Vi). These languages belong to diverse language families and employ distinct writing systems, thereby introducing additional challenges for low-resource machine translation. This paper assesses the performance of several state-of-the-art MNMT models in combination with recent fine-tuning methods. The main contributions of this paper can be summarized as follows:

- Proposing a compact and language-sensitive MNMT framework to reduce the number of trainable parameters;
- Conduct a comprehensive evaluation of various MNMT models on selected Asian languages under low-resource settings;
- Analyzing linguistic factors that may influence MNMT performance.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed MNMT models. Section 4 presents and discusses the experimental results. Section 5 concludes the paper and outlines directions for future research.

2. Methods

This section briefly introduces neural machine translation (NMT) as a foundational background, followed by a description of Google's MNMT and the proposed Compact and Language-Sensitive MNMT models for low-resource settings. It also provides an overview of multi-stage training, which represents an effective strategy for organizing training data to better exploit multilingual signals.

2.1. Neural Machine Translation

Given a set of sentence pairs $D = \{(x, y)\}$, the encoder f_{enc} with parameters θ_{enc} maps an input sequence $x = (x_1, x_2, \dots, x_n)$ to a sequence of continuous representations $h^{enc} = (h_1^{enc}, h_2^{enc}, \dots, h_n^{enc})$, whose length depends on the source sentence. The decoder f_{dec}, parameterized by θ_{dec}, then generates an output sequence $y = (y_1, y_2, \dots, y_m)$ by modeling the conditional probability $P(y_t)$ as follows:

$$P(y_t) = softmax\big(f_{dec}(h_{dec}, c_t)\big) \tag{1}$$

where h^{dec} denotes the sequence of continuous representations produced by the decoder, and c_t is the context vector, which is computed as follows:

$$c_t = \sum_{i=1}^{n} a_{t,i} h_i^{enc} \tag{2}$$

where $a_{t,i}$ is attention weight:

$$a_{t,i} = softmax(e_{t,i}) \tag{3}$$

where $e_{t,i}$ denotes the similarity score between the source and target representations. The parameters used to compute the cross-attention weights $a_{t,i}$ are denoted by θ_{attn}. The $e_{t,i}$ can be computed as described in [2]:

$$e_{t,i} = V_\alpha^T tanh(W_{dec} s_{t-1}^{dec} + W_{enc} h_i^{enc}) \tag{4}$$

where W_{dec}, W_{enc} are learnable parameters of the attention layer.

Subsequently, the target hidden state s_t is updated as follows:

$$s_t = f_{enc}(s_{t-1}, y_{t-1}, c_t) \tag{5}$$

The encoder and decoder are jointly trained to maximize the conditional probability of the target sequence given the source sequence:

$$L_t(D;\theta) = \sum_{d=1}^{|D|} \sum_{t=1}^{M} logP(y_t, x; \theta_{enc}, \theta_{dec}, \theta_{attn}) \tag{6}$$

where M denotes the length of the target sentence.

Both the encoder and decoder can be implemented using different fundamental neural architectures, including Recurrent Neural Networks (RNNs) [4], [5], Convolutional Neural Networks (CNNs) [5], and self-attention-based models [6].

2.2. Multilingual Neural Machine Translation

2.2.1. Google's MNMT

This idea was originally proposed by [12], who introduced a multilingual machine translation model based on a single unified architecture capable of handling multiple language pairs simultaneously, rather than training a separate model for each pair. This design enables the model to learn shared representations across languages, allowing knowledge transfer among language pairs and facilitating mutual reinforcement during training.

In addition to training multiple language pairs within a single model, the authors proposed a simple yet effective data modification strategy: prepending a special token to the beginning of each source sentence to indicate the desired target language. This token explicitly specifies the translation direction and guides the model during decoding. For example, when translating into a target language denoted as "zz" a token such as "2zz" is added to the source sentence to signal the intended target language. The objective function now becomes:

$$L_t(D;\theta) = \sum_{l=1}^{L} \sum_{d=1}^{|D|} \sum_{t=1}^{M} logP(y_t, x; \theta_{enc}, \theta_{dec}, \theta_{attn}) \tag{7}$$

with L is the number of language pairs.

2.2.2. Compact & Language-Sensitive MNMT

A Compact and Language-Sensitive model [13] is proposed by using an attention-based architecture that reduces the number of parameters in the original Transformer model [7]. The model is specifically designed for multilingual translation, aiming to improve parameter efficiency while enhancing language-specific adaptability within a unified framework.

JTE HCM-UTE
ISSN: 2615-9740
JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Engineering
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

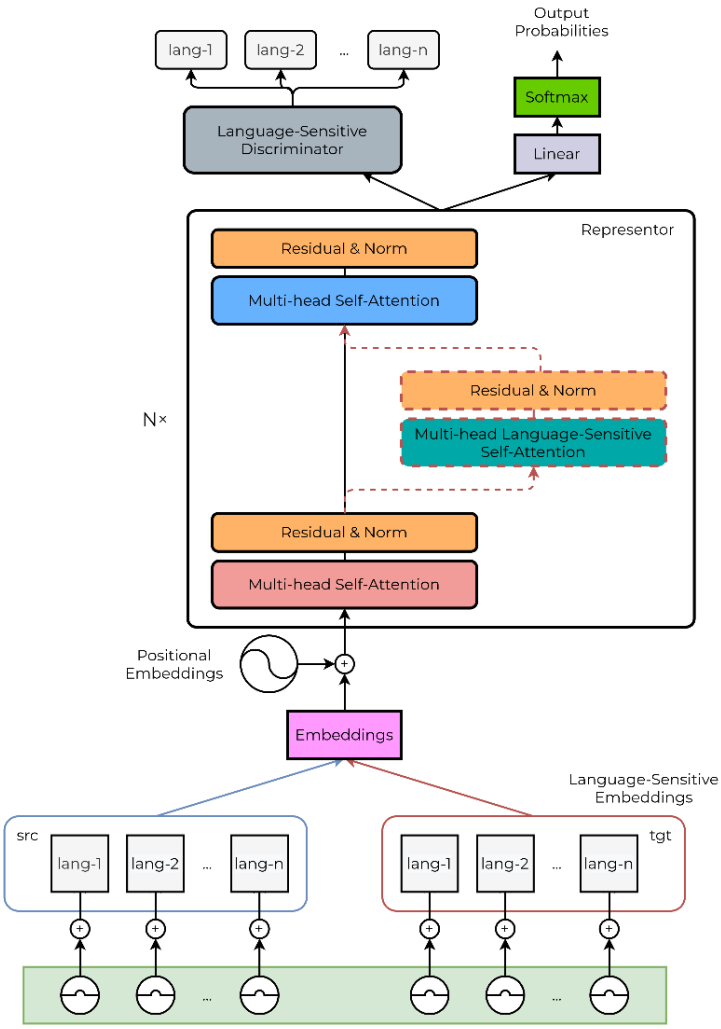

Figure 1. *Compact and Language-sensitive model.*

1) Compact representation

In neural machine translation architectures, the encoder and decoder constitute two fundamental components that perform analogous functions and exhibit similar layer-wise structures. It is observed that these components can share a common set of parameters. Accordingly, the proposed model adopts a unified representation that replaces separate encoder and decoder modules by sharing the parameters of the self-attention layers, feed-forward networks, and their associated normalization components, as illustrated in Figure 1. The parameter for the representation set is denoted as θ_{rep}. Therefore, the training target function is changed to:

$$L_{m-t}(D;\theta) = \sum_{l=1}^{L} \sum_{d=1}^{|D_l|} \sum_{t=1}^{M} logP\left(y_t^l \vee x^l, y_t^l; \theta_{rep}, \theta_{attn}\right) \quad (8)$$

Language sensitive modules. Besides the word embedding and positional embedding as in the Transformer architecture, the authors propose to add a language embedding to express clearly the language that the sentence belongs to instead of just putting extra special characters as in the basic translation model [12]. The operation is similar to that of the embedding set in Transformer architecture, embedding values from the language embeddings are added directly to the representation of each word of the input sentence, this helps the model to identify the language. This embedding set is denoted as $E_{lang} \in R^{|K|*d_{model}}$, where $|K|$ denotes the number of languages involved, and d_{model} represents the hidden dimensionality.

Language sensitive attention. In NMT architecture, cross-attention only occurs in the decoder to identify the most important parts of the source sentence to generate the output in target language. For multilingual machine translation, the authors introduced 3 different ways to design a cross-attention mechanism, including i) shared attention, ii) mixed attention iii) linguistic attention used in the proposed method.

i) Shared attention: uses both cross-attention and self-attention.
ii) Mixed attention: this mechanism uses its own cross-attention with self-attention, but is used for all language pairs (similar to the basic multilingual translation model).
iii) Language sensitive attention: allows the model to actively select parameter sets of cross-attention depending on language pairs.

It has been indicated that both shared and mixed attention mechanisms may struggle to effectively extract information from different source language pairs when decoding multiple source languages with divergent word orders. For example, Vietnamese follows a Subject-Verb-Object (SVO) structure, whereas Japanese employs a Subject-Object-Verb (SOV) word order, leading to potential confusion during attention-based decoding. Therefore, the compact and sensitive model mainly focuses on the language sensitive attention in this model. To accomplish this, the author uses multiple parameter sets θ_{attn} to represent different translation pairs. However, language-sensitive attention does not support zero-shot translation, as no parameter set is available for unseen language pairs. Consequently, to enable zero-shot translation, a mixed attention mechanism must be adopted.

2) Language sensitive modules

The compact representation promotes extensive parameter sharing and leverages similarities across languages. Nevertheless, this approach diminishes the model's capacity to discriminate among distinct languages. To mitigate this issue, the authors propose the incorporation of three additional language-aware modules, which are detailed in the following modules.

Language-sensitive Discriminator. With this new model, the representation set used for both encoders and decoders will take advantage of the similarities of the languages, but weaken the model's ability to distinguish between different languages (this problem can cause the model to not translate sentences correctly into the language they need to actually translate). Since then the authors have introduced an additional module, the Language-sensitive Discriminator, to reinforce the ability to distinguish languages in model representations.

In machine translation models, the hidden state of the last layer can be considered an abstract representation of the sentence being translated. For this language-sensitive discriminator module, this paper deploys a neural network f_{dis} to the top-layer representation h_{top}^{rep}, and the model outputs a language classification distribution P_{class}:

$$h^{dis} = f_{dis}\left(h_{top}^{rep}\right)$$
$$P_{class}(d) = softmax\left(W_{dis} * h_d^{dis} + b_{dis}\right) \quad (9)$$

with $P_{class}(d)$ is a score for language assessment of a sentence pair d, W_{dis}, b_{dis} or θ_{dis} denote parameters. This score P_{lang} is used for lingual distinction of h_d^{dis}. The authors proposed two types of different neural networks for f_{dis}, including a convolutional network with max pooling and a feed forward network. Results, f_{dis}, the two types of neural networks are identical, so this paper adapts a feed forward network for simplicity.

A objective of discrimination function as follow:

$$L_{dis}(\theta_{dis}) = \sum_{k \in K} \sum_{d=1}^{|K|} I\{g_d = k\} * logP_{lang}(d) \quad (10)$$

with $I\{\cdot\}$ denotes the indicator function, and g_d corresponds to language k.

Finally, once the language-sensitive discriminator is integrated, the model is trained jointly on all language pairs D using the objective function defined below:

$$L(D;\theta) = L\left(D;\theta_{rep},\theta_{attn},\theta_{dis}\right) = (1-\lambda)L_{m-t}\left(\theta_{rep},\theta_{attn}\right) + \lambda L_{dis}(\theta_{dis}) \quad (11)$$

with λ denotes a learn-able parameter or already identified in order to balance the priority between translation and language evaluation. In the experiments, $\lambda = 0.05$ gave the best performance.

2.2.3. Multistage Fine-Tuning for NMT

Unlike previous work of NMT [12] which only adopted single phase training for multilingual translation for multiple language pairs, this new method [15] comes with the purpose of improving performance of NMT models for translating from English to other low-resource languages based on exploring multi-lingual properties through multi-phase transfer learning (pre-train, fine-tuning). Additionally, this method mainly focus on making the optimum use of a large-scale dataset of English and another language (such as France) combined with a multi-directional parallel corpus, i.e the same content in many different languages to apply in one-to-many translation model to exploit the multilingualism to improve the quality of English translation pairs into low-resource languages.

The proposed model focuses on translating from English into N different languages. Specifically, this paper utilizes two types of corpora:

- Small-scale multi-directional parallel corpus En-YY_1-...-YY_N, comprises English and N target languages need to be improved translation quality.
- Medium-scale multi-directional parallel corpus, En-XX, with XX denotes only supported target language and it is not necessary one of the target language in the multi-directional parallel corpus, $YY_k (1 \le k \le N)$. In addition, corpus of this language pair only used in the pre-training phase for knowledge initialization in the pre-processing phase and fine-tune in phase 2, do not have any roles in the last phase.

The multistage training method could be depicted in Figure 2 as follow:

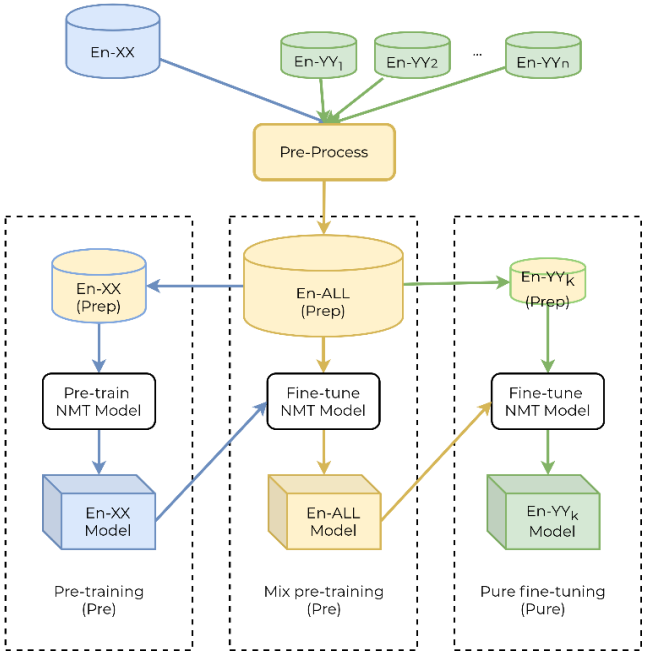

Figure 2. *Multistage Fine-Tuning for NMT.*

- **Pre-training (Pre):** employ a single NMT model trained on a corpus from a high-resource language pair, En-XX. This step of training aims to take advantage of the parallel corpus to initialize a strong encoder for English and use it for the next fine-tuning steps.

- **Mixed pre-training / Mixed fine-tuning (Mix):** Model training was conducted either from initialization or via continued training, using a combined dataset consisting of the En-XX parallel corpus and one or more low-resource En-YY language pairs. At this stage, it is time for multi-language corpus to show its strength, with the simultaneous training of multiple language pairs with the same source sentence as English while having a strong encoder from pre-training the main task of this stage is to focus on learning a decoder for all language pairs. This phase does not eliminate the original large dataset and is mixed with the multi-directional parallel dataset to train, which aims to utilize the large corpus as well as create a "smooth" transition between the corpus.
- **Pure fine-tuning (Pure):** Training then proceeds using only the corpus corresponding to the specific En-YY language pair. This stage concentrates the knowledge acquired in the previous steps to further adapt the model to the target language pair.

3. Results and Discussion

This section presents details of the datasets, NMT configurations, and evaluation metrics used in all experiments.

3.1. Datasets

The following corpora in Table 1 are used in all experiments:

- An English-Japanese parallel corpus from the Kyoto Free Translation Task (KFTT) is used. This corpus was originally developed by the National Institute of Information and Communications Technology and released as the Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles [16].
- A multilingual parallel corpus from the Asian Language Treebank (ALT) is employed. This corpus comprises multiple languages, including English (En), Bengali (Bn), Filipino (Tl), Indonesian (Id), Japanese (Ja), Khmer (Km), Malay (Ms), and Vietnamese (Vi). The construction of ALT began with the selection of approximately 20,000 sentences from English Wikinews, which were subsequently translated into the remaining languages [17].

Table 1. *Statistics of parallel corpora*

Dataset	Training set	Valid set	Test set
KFTT En-Ja	440,288	1,166	1,160
ALT En-{Bn,Tl,Id,Ja,Km,Ms,Vi}	18,088	1,000	1,018

3.2. Set-up and Configurations

This sub-section provides common as well as different configurations for the translation models: NMT, Google's MNT, and Compact & Language-sensitive MNMT. For the NMT baseline and Google's MNT, the paper uses the Transformer model implemented in the Fairseq framework[1]. For the Compact & Language-sensitive MNMT, the paper also implements it based on the Fairseq code.

Common **configurations.** The paper trains all models using the default Transformer configuration, consisting of a 6-layer encoder and a 6-layer decoder with 512-dimensional hidden representations. The number of head is $h = 8$. Dropout rates, for each stage, are $P_{drop1} = 0.2$, $P_{drop2} = 0.3$, and $P_{drop3} = 0.1$, respectively. Each mini-batch contains approximately 3,000 source and 3,000 target tokens from a single translation direction. The Adam optimizer [18] is with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$, and $warmup_{steps} = 4000$. Label smoothing is applied with a smoothing factor of $\epsilon_{ls} = 0.1$. During evaluation, beam search is employed with a beam size of $k = 4$ and a length penalty of $\alpha = 0.6$.

[1] https://github.com/facebookresearch/fairseq

Different configurations. While the above configurations are enough for NMT and Google's MNMT, the Compact & Language-sensitive MNMT needs some minor changes including a 6-layer representor and the discriminator embedding size of 1024.

Translation quality is evaluated by using case-insensitive BLEU scores [19]. In addition, statistical significance of BLEU score differences between the proposed models and the baseline systems is assessed using the bootstrap resampling method [20], with a significance threshold of $p < 0.05$.

3.3. Main results

Table 2 shows that the baseline single-language NMT model performs worst across all languages, confirming its limitations in low-resource and diverse linguistic settings. Google's MNMT achieves strong improvements via large-scale multilingual transfer, providing the highest scores for several languages, especially Bn, Km, Ms, and Vi. Notably, the proposed Compact & Language-Sensitive MNMT performs competitively despite its smaller size, surpassing Google's system on most languages (Tl, Id, Ja, Km, Ms) and matching it closely elsewhere. These results demonstrate that language-sensitive training yields highly efficient cross-lingual sharing, enabling a compact model to rival or exceed a much larger MNMT system. Overall, the table highlights the clear advantage of multilingual modeling over single-language NMT and shows that the proposed approach achieves state-of-the-art performance with far fewer parameters.

Table 2. *Experiments on NMT, Google's MNMT, and Compact & Language-sensitive MNMT. The XX column denotes the high-resource external parallel corpus (Ja). The columns corresponding to the YY test sets represent the target languages in the ALT multilingual corpus, with the best scores for each target language highlighted in bold, and* ***** *marks the scores which are better than the ones of the single NMT model.*

Model	XX	Training			YY test set						
		Pre	Mix	Pure	Bn	Tl	Id	Ja	Km	Ms	Vi
#1. NMT	-	-	-	Yes	3.2	22.5	25.6	10.5	21.4	30.7	26.8
#2. Google's MNMT	Ja	Yes	Yes	Yes	**10.8***	28.6*	28.9*	22.6*	30.0*	34.8*	37.0*
#3. Compact & Language-sensitive MNMT	Ja	Yes	Yes	Yes	9.5*	**30.6***	**33.8***	**24.2***	**31.0***	**38.6***	36.4*

Discussion on experiment results. Table 2 indicates that the combination of MNMT and Multi-stage training method yields better performance compared to the single NMT. However, do these results indicate that MNMT really takes advantage of extra knowledge from other languages?

The ALT corpus is a multilingual parallel dataset comprising English, Bengali, Filipino, Indonesian, Japanese, Khmer, Malay, and Vietnamese. It was constructed by first selecting approximately 20,000 sentences from English Wikinews, which were subsequently translated into the remaining languages. Although the corpus includes seven target languages, it does not introduce additional source content; rather, it provides aligned translations across multiple languages for the same set of sentences.

As shown in Table 2, most language pairs exhibit comparable improvements, with the exception of the English-Japanese (En-Ja) pair, which achieves substantially superior performance owing to the availability of a large bilingual corpus (KFTT En-Ja). These results suggest that, when multiple language pairs are trained jointly, semantically equivalent sentences in different target languages tend to converge toward similar representations, thereby complementing one another and improving overall translation quality within the system. This finding highlights the effectiveness of multilingual learning in MNMT models, even without relying on large-scale source corpora. The effect is particularly notable given the relatively small size of the multi-parallel dataset (approximately 18,000 translated sentences per language), which is modest compared to high-resource bilingual corpora.

The empirical findings demonstrate the considerable value of multi-parallel datasets, despite their current scarcity. Moreover, constructing multi-parallel data may be more cost-effective than building

multiple large bilingual corpora, as adding a new language under the ALT framework simultaneously creates parallel data with all existing languages. This characteristic underscores the promising potential of multi-parallel datasets for advancing multilingual machine translation research.

Discussion on language properties. Table 3 conducts a survey of languages in the Asian Treebank Language (ALT) dataset, comparing the similarities between these languages. From there, the paper gives comments on the results obtained in the empirical results.

Table 3. *Statistical information about target languages in the ALT dataset.*

Code	Language	Language Family	Word Order	Typology
Ja	Japanese (Japanese)	Japonic	SOV	Agglutinative
Id	Indonesian (Indonesian)	Austronesian	SVO	Agglutinative
Ms	Malay (Malay)	Austronesian	SVO	Agglutinative
Tl	Filipino (Tagalog)	Austronesian	OSV or SVO	Agglutinative
Km	Cambodian (Khmer)	Austro-Asiatic	SVO	Agglutinative
Vi	Vietnamese (Vietnamese)	Indo-European	SVO	Isolating
Bn	Bangladeshi (Bengali)	Indo-European	SVO	Flexional

About the Compact & Language-sensitive MNMT, this model mainly focused on how to train pure multilingual translation (i.e. training on many consecutive language pairs). The experiments mentioned an observation during the multi-stage training for the Compact & Language-sensitive MNMT model that it works very well at the stage of mixed refinement (i.e, the part of pure multilingual training) but at the end (monolingual training), the results do not improve much. From these observations, the paper draws a remark that the higher the similarity of language pairs, the better the results will be. This explains the comparison results with the original multi-stage training method in Table 2 as follows:

- En-Ja pair: Despite substantial linguistic differences and divergent word order compared to the other languages, the English-Japanese (En-Ja) pair still achieves a +1.6 improvement, indicating that it benefits from knowledge transferred from the remaining languages within the multilingual training framework.
- En-Ms and En-Id pairs: The English-Malay and English-Indonesian pairs achieve notable improvements of +3.8 and +4.9, respectively. These gains likely stem from their close linguistic relatedness, including shared language family, typology, and word order, which facilitates more effective knowledge transfer.
- En-Tl pair: The English-Filipino pair shows a relatively strong improvement of +2.0. Although Filipino belongs to the same Austronesian language family as Malay and Indonesian and shares certain typological features, differences in word order may limit the extent of knowledge transfer, resulting in comparatively smaller gains.
- En-Km and En-Vi pairs: The English-Khmer and English-Vietnamese pairs show performance comparable to the baseline multi-stage training method, with gains of +1.0 and −0.6, respectively. Although Khmer and Vietnamese share certain typological and word-order characteristics, they are linguistically more distant from the other languages included in the training set, which may limit effective knowledge transfer and result in smaller improvements compared to the En-Ms, En-Id, and En-Tl pairs. Furthermore, Vietnamese, as an Austroasiatic language with distinctive structural properties, differs typologically from most of the remaining languages in the model, which may contribute to its slight performance degradation relative to the En-Km pair.
- En-Bn pair: The English-Bengali pair is the only language pair that performs substantially worse than the baseline training method, with a decrease of −1.3. This result may be attributed to its pronounced grammatical and typological differences from the other language pairs in the ALT set, which limit effective cross-lingual knowledge transfer. In addition, the baseline bilingual

model for this pair already exhibits relatively low performance (BLEU 3.2), suggesting limited translation capacity. Consequently, the En-Bn pair benefits less from multilingual training compared to other language pairs that share greater linguistic commonalities.

Discussion of the parameter counts. The standard Transformer contains approximately 92M trainable parameters, while the Compact & Language-Sensitive MNMT model uses only ~42M, representing a reduction of about 54%. This substantial decrease highlights the parameter efficiency of the proposed architecture. With less than half the parameters, the compact model reduces memory and computational requirements, making it more suitable for scalable and resource-constrained multilingual settings while maintaining effective modeling capacity.

4. Conclusions

This paper presents a systematic empirical investigation of the effectiveness of state-of-the-art multilingual neural machine translation (MNMT) models for low-resource language pairs with limited linguistic relatedness, including Bengali, Filipino, Indonesian, Japanese, Khmer, Malay, and Vietnamese. The study begins with an analysis of the linguistic similarities and differences among these languages, followed by an evaluation of various MNMT models and a detailed examination of the resulting translation outputs. Based on the experimental findings, several conclusions are drawn that may inform and facilitate future research on machine translation for Asian languages.

Future work will extend this study by incorporating a broader range of languages and conducting experiments with more advanced model architectures and more comprehensive analytical frameworks. In addition, MNMT may be combined with complementary approaches, such as unsupervised learning, back-translation, and grammar structure [21], to further enhance translation performance for low-resource language pairs.

Acknowledgments

The authors would like to thank the reviewers for their valuable comments.

Conflict of Interest

The authors declare no conflict of interest.

REFERENCES

[1] R. Dabre, T. Nakagawa, and H. Kazawa, "A survey of multilingual neural machine translation," *ACM Comput. Surv.*, vol. 53, no. 5, pp. 1–38, 2020.
[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Representations (ICLR)*, San Diego, CA, USA, 2015.
[3] M. Ü. Uyar, "RNNs with attention for machine translation," in *Machine Learning and AI with Simple Python and Matlab Scripts: Courseware for Non-computing Majors*. Hoboken, NJ, USA: IEEE-Wiley, 2025, pp. 209–223, doi: 10.1002/9781394294985.ch13.
[4] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 27, 2014, pp. 3104–3112.
[5] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," *arXiv preprint*, arXiv:1409.1259, 2014.
[6] A. Li, "Application of convolution neural network algorithm in English translation," in *Proc. Int. Conf. Integrated Intelligence and Communication Systems (ICIICS)*, Kalaburagi, India, 2023, pp. 1–8, doi: 10.1109/ICIICS59993.2023.10421570.
[7] A. Vaswani *et al*., "Attention is all you need," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5998–6008.
[8] Y. Chen, Y. Liu, Y. Cheng, and V. O. K. Li, "A teacher–student framework for zero-resource neural machine translation," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Vancouver, BC, Canada, 2017, pp. 1925–1935, doi: 10.18653/v1/P17-1176.
[9] Y. Chen, Y. Liu, and V. O. K. Li, "Zero-resource neural machine translation with multi-agent communication game," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5086–5093.
[10] Y. Cheng, Q. Yang, Y. Liu, M. Sun, and W. Xu, "Joint training for pivot-based neural machine translation," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Melbourne, VIC, Australia, 2017, pp. 3974–3980, doi: 10.24963/ijcai.2017/555.
[11] O. Firat, K. Cho, and Y. Bengio, "Multi-way, multilingual neural machine translation with a shared attention mechanism," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT)*, San Diego, CA, USA, 2016, pp. 866–875, doi: 10.18653/v1/N16-1101.
[12] M. Johnson *et al*., "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 339–351, 2017.
[13] Y. Wang *et al*., "A compact and language-sensitive multilingual translation method," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Florence, Italy, 2019, pp. 1213–1223.
[14] B. Zoph and K. Knight, "Multi-source neural translation," in *Proc. NAACL-HLT*, San Diego, CA, USA, 2016, pp. 30–34, doi: 10.18653/v1/N16-1004.
[15] R. Dabre and A. Fujita, "Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Nov. 2019, pp. 1410–1416.

[16] G. Neubig, "The Kyoto free translation task," 2011. [Online]. Available: http://www.phontron.com/kftt
[17] H. Riza *et al*., "Introduction of the Asian language treebank," in *Proc. Oriental COCOSDA*, 2016.
[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2015.
[19] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Philadelphia, PA, USA, 2002.
[20] P. Koehn, "Statistical significance tests for machine translation evaluation," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, Spain, Jul. 2004.
[21] H. Nguyen *et al*., "Language-oriented sentiment analysis based on grammar structure and improved self-attention network," in *Proc. 15th Int. Conf. Evaluation of Novel Approaches to Software Engineering (ENASE)*, Prague, Czech Republic, May 2020, pp. 339–346.

Hong Buu Long Nguyen received the B.S. degree (Honors Program) in Information Technology from the University of Science, Vietnam National University, Ho Chi Minh City (VNU-HCM), Vietnam, in 2010, the M.S. degree in Computer Science from the same university in 2015, and the Ph.D. degree in Computer Science in 2023. He is currently a lecturer and researcher at the University of Science, Vietnam National University, Ho Chi Minh City. He has published numerous articles in reputable journals and serves as a reviewer for several A*, A, and B-ranked conferences as well as SCIE-indexed journals. His research interests include machine translation, question answering, and language modelling.

Email: nhblong@fit.hcmus.edu.vn. ORCID: https://orcid.org/0000-0002-0884-1635

Thanh Tung Vu is a graduate student in the Faculty of Information Technology, University of Science, VNU-HCMC., Vietnam. His research interests are Natural Language Processing, Deep Learning and Machine Translation.

Email: thanhtungvu727@gmail.com. ORCID: https://orcid.org/0009-0000-2837-3288