https://doi.org/10.31449/inf.v47i3.4761 Informatica **47** (2023) 349–360 **349**

Khmer-Vietnamese Neural Machine Translation Improvement Using Data Augmentation Strategies

Thai Nguyen Quoc[1], Huong Le Thanh[1,*] and Hanh Pham Van[2]
[1]School of Information and Communication Technology, Hanoi University of Science and Technology, Vietnam
[2]FPT AI Center
E-mail: thai.nq212642m@sis.hust.edu.vn, huonglt@soict.hust.edu.vn, hanhpv@fsoft.com.vn

Keywords: machine translation, data augmentation, low-resource, Khmer-Vietnamese

Received: March 22, 2023

The development of neural models has greatly improved the performance of machine translation, but these methods require large-scale parallel data, which can be difficult to obtain for low-resource language pairs. To address this issue, this research employs a pre-trained multilingual model and fine-tunes it by using a small bilingual dataset. Additionally, two data-augmentation strategies are proposed to generate new training data: (i) back-translation with the dataset from the source language; (ii) data augmentation via the English pivot language. The proposed approach is applied to the Khmer-Vietnamese machine translation. Experimental results show that our proposed approach outperforms the Google Translator model by 5.3% in terms of BLEU score on a test set of 2,000 Khmer-Vietnamese sentence pairs.

Povzetek: Raziskava uporablja predhodno usposobljen večjezični model in povečanje podatkov. Rezultati presegajo Google Translator za 5,3%.

1 Introduction

Machine translation (MT) is the task of automatically translating text from one language to another. There are three common approaches to MT: rule-based approach [1], statistical-based approach [2, 3], and neural-based one [4, 5, 6]. The rule-based approach depends on translation rules and dictionaries created by human experts. Statistical Machine Translation (SMT) relies on techniques like word alignment and language modeling to optimize the translation process. While SMT can handle a wide range of languages and translation scenarios, it often struggles with capturing complex linguistic phenomena and handling long-range dependencies. With significant advancements in deep learning, Neural Machine Translation (NMT) approaches have shown great potential and have replaced SMT as the primary approach to MT. NMT models capture contextual information, handle word reordering, and generate fluent and natural translations. NMT has gained popularity due to its end-to-end learning, ability to handle complex linguistic phenomena, and improved translation quality. Among all NMT systems, transformer-based MT models [7, 8] have demonstrated superior performance. The key feature of transformer models [8] is their attention mechanism, which allows them to effectively capture dependencies between different words in a sentence. Unlike traditional recurrent neural networks that process words sequentially, transformers can consider the entire input sentence simultaneously. This parallelization significantly speeds up the training process and makes transformers more efficient for long-range dependencies.

One notable limitation of NMT techniques pertains to their reliance on a substantial number of parallel sentence pairs to facilitate model training. Unfortunately, most of language pairs in the world are lack of such a large dataset. Consequently, these language pairs fall under the category of low-resource MT, presenting a challenging scenario for the application of neural-based models in this domain.

Several works have carried out research to solve the low-resource problem in NMT. Chen et al.[9], Kim et al. [10] dealt with the low-resource NMT by using pivot translations, where one or more pivot languages were selected as a bridge between the source and target languages. The source-pivot and the pivot-target should be rich-resource language pairs. Sennric et al. [11], Zhang [12] applied the forward/backward translation approaches to generate parallel sentence pairs by translating the monolingual sentences to the target/source language via a translation system. Then, the pseudo parallel data was mixed with the original parallel data to train an NMT model. A problem in this approach is how to control the quality of the pseudo parallel dataset in order to improve the performance of the low-resource NMT system.

Since NMT requires the capability of both language understanding (e.g., NMT encoder) and generation (e.g., NMT decoder), pre-training language model can be very helpful for NMT, especially low-resource NMT. To do this task, BART model [13] has been proposed to add noises and randomly masked some tokens in the input sentences in the encoder, and learn to reconstruct the original text in the decoder. T5 model [14] randomly masks some tokens and replace the consecutive tokens with a single sentinel

token.

To address the low-resource problem in NMT, we propose to fine-tune mBART [15] - a pretrained multilingual Bidirectional and Auto-Regressive Transformers model that has been specifically designed for multilingual applications, including MT. The fine-tuning process is combined with several strategies, including the utilization of back-translation techniques [11] and data augmentation via a pivot language. We propose several data augmentation strategies to augment training data as well as controlling the data quality.

Our proposed approach can be applied to any low-resource language pairs. However, in this research, we evaluate our approach by implementing it with the low-resource Khmer-Vietnamese (Km-Vi) language pair, using a dataset with 142,000 parallel sentence pairs from Nguyen et al. [16]. As far as we know, there is only two works dealing with the Km-Vi machine translation ([17], [18]). Nguyen et al. [17] presented an open-source MT toolkit for low-resource language pairs. However, this approach only used a transformer architecture to train the original dataset, without applying fine-tuning, transfer learning, or additional data augmentation techniques. Pham and Le [18] fine-tuned mBART and applied some data augmentation strategies. In this research, we have extended the work in [18] to improve the performance of the Km-Vi NMT system. The contributions are as follows:

- We propose new methods for data selection based on sentence-level cosine similarity through the bi-encoder model [19] combined with the TF-IDF score.
- We suggest a data generation strategy to generate best candidates for the synthetic parallel dataset.
- To control the quality of augmented data, we propose an "aligned" version to enrich the data and a two-step filtering to eliminate low quality parallel sentence pairs.

The remainder of this paper is organized as follows. Section 2 analyzes various techniques in existing research to address the limitations of low-resource NMT. Section 3 describes our system diagram. Our proposed data augmentation strategies are outlined in Section 4. Section 5 elaborates on the experimental design, whereas Section 6 presents an analysis of the empirical outcomes. Finally, Section 7 concludes our paper.

2 Related work

Pretrain Language Models (PLMs) have proven to be helpful instruments in the context of low-resource NMT. Literature has shown that low-resource NMT models can benefit from the use of a single PLM [20, 21] or a multilingual one [13]. The multilingual PLM is claimed to facilitate more effective learning of the connection between the source and the target representations for translation. These transfer learning methods leverage rich-resource language pairs to train the system, then fine-tune all the parameters on the specific target language pair [22]. The rich-resource language pairs should be in a similar language family to the low-resource ones, to have good results.

Data augmentation is the method of generating additional data, achieved by expanding the original dataset or integrating supplementary data from relevant sources. Various approaches to data augmentation have been explored, including: (i) paraphrasing and sentence simplification [23], (ii) word substitution and deletion [24, 25], (iii) limited and constrained word order permutation [26], (iv) domain adaptation [27], (v) back-translation [11], and (vi) data augmentation via a pivot language [28].

Paraphrasing and sentence simplification [23] offer varied quality, with a risk of introducing semantic changes or losing important information. Word substitution [24] requires careful selection of synonyms to maintain accuracy, while word deletion [25] can introduce noise and requires effective training to handle missing information. Limited and constrained word order permutation [26] suits language pairs with word order variations but requires defining complex constraints based on language characteristics. Domain adaptation [27] addresses the challenge of domain-specific low-resource machine translation, which is not the target of this research. Back-translation [11] has proven successful by generating synthetic source sentences through translating target sentences. However, this approach carries a risk of errors due to imperfections in pre-trained translation models. On the other hand, the pivot-based approach [28] involves translating low-resource language pairs through a high-resource language. This approach relies on good translation quality to and from the pivot language.

Back-translation and pivot-based translation are considered reliable and generalizable approaches when complemented by effective post-processing methods for filtering low-quality data. Therefore, this paper specifically concentrates on utilizing back-translation and pivot-based translation as the selected methods for data augmentation. To improve the quality of the synthetic parallel data generated by these methods, two strategies are employed: (i) data selection and (ii) synthetic data filtering.

Data selection is the process of ranking and selecting a subset from a target monolingual dataset that ensures in-domain as the training data. The objective of this process is to improve the performance of an NMT system for a particular domain. Various techniques for data selection have been proposed in the literature, such as computing sentence scores based on Cross-Entropy Difference (CED) [29, 30], and using representation vectors to rank sentences in the monolingual dataset [31, 32]. Three data selection methods had been implemented by Silva et al. [33], namely CED, TF-IDF, and Feature Decay Algorithms (FDA) [34]. The experimental results pointed out that the TF-IDF method gained the best improvements in both BLEU and TER (Translation Error Rate) scores.

Synthetic data filtering To filter out low-quality sen-

tence pairs, Imankulova et al. [35] proposed a method based on the BLEU measure. This method involves leveraging a source-to-target NMT model to translate the synthetic source sentences into synthetic target sentences. Subsequently, the sentence-level BLEU score is calculated for each sentence pair between the synthetic target sentence and the target sentence, with the ultimate objective of excluding low-score sentences. Koehn et al. [36] proposed another approach based on the sentence-level cosine similarity of two sentences. However, their proposal required an effective acquisition of the linear mapping relationship between the two embedding spaces of the source language and the target one.

Another way to improve translation quality is by using data augmentation methods via a pivot language [28]. This method involves translating sentences from the source language to the pivot one using the source-pivot translation model, followed by translating sentences in the pivot language to sentences in the target language. However, there are certain restrictions associated with this technique. Firstly, the circular translation process increases the decoding time during inference as it can iterate through multiple languages to obtain the desired quality. Secondly, translation errors may arise in each step, which can lead to low-quality translation of the sentence in the target language.

In this paper, we introduce an approach aimed at enhancing the performance of low-resource MT. Our approach incorporates multiple data augmentation strategies alongside various data filtering methods to improve the quality of synthetic data. In the subsequent sections, we introduce these methods in detail.

3 Our system diagram

As previously mentioned, our goal is to propose strategies that can improve the performance of low-resource NMT systems. The proposed approach will be applied for the Km-Vi language pair. To do this, we first fine-tune the mBART50 [37] model with the Km-Vi bilingual dataset.

The mBART model Multilingual BART (mBART) [15] is a sequence-to-sequence denoising auto-encoder that was pre-trained on large-scale monolingual corpora in many languages using the BART objective [13]. The pre-trained task is to reconstruct the original text from the noise one, using two types of noise: random span masking and order permutation. A special variant of mBART called mBART50 [37], has been trained in 50 languages, including Khmer and Vietnamese. Nonetheless, the mBART50's translation quality of the Km-Vi language pairs is low. To deal with this problem, we propose to fine-tune the mBART50 with the Km-Vi bilingual dataset combined with the augmented dataset through several strategies.

Our proposed Khmer-Vietnamese MT system model is described in Figure 1, which incorporates two strategies for data augmentation: (i) back-translation with a dataset in the target language; and (ii) data augmentation via English pivot language. These strategies will be introduced in the next section.

4 Data augmentation strategies

Since the word orders and theirs meaning in machine translation are important, methods such as paraphrasing, simplification, limited and constrained word order permutation cannot provide good parallel sentence pairs.

4.1 Back-translation with a dataset in the target language

Back-translation method proposed by Senrich et al [11] is an useful way to generate additional training data for low-resource NMT. This method leverages an external dataset in the target language, termed the "target-language dataset". It employs a target-to-source NMT model, trained on the original bilingual dataset, to translate this dataset into the source language. The resulting translated sentences are then combined with their corresponding target sentences, creating a synthetic bilingual dataset. However, the dataset's quality generated by this method is not guaranteed. To address this issue, we improve this method by integrating data filtering techniques to the back-translation process. Our proposed method is conducted in three steps as follow:

- **Step 1 - Data selection:** Rank and select sentences from a target-language dataset that is in the same domain as sentences in the original bilingual dataset.
- **Step 2 - Data generation:** Each sentence from the output dataset in Step 1 is translated to k sentential candidates in the source language using the target-to-source NMT model which has been trained by fine-tuning the mBART50 with the original bilingual dataset.
- **Step 3 - Data filtering:** Filter out low-quality bilingual sentence pairs in the synthetic parallel dataset.

We will discuss these three steps in the following sections.

4.1.1 Data selection

For a given dataset D consisting of T sentence pairs in a specific domain, and a set of sentences in a general domain G, the aim of data selection is to rank the sentences in G based on their similarity to the domain of D, then selecting highest-ranked sentences to form a subset that shares the same domain as D. Given that TF-IDF is a popular technique used to identify representative words for a dataset, we can assess whether sentences in G belong to the same domain as D using this measure. In addition to the TF-IDF measure, cosine similarity can be employed to measure the semantic similarity between two sentences based on their

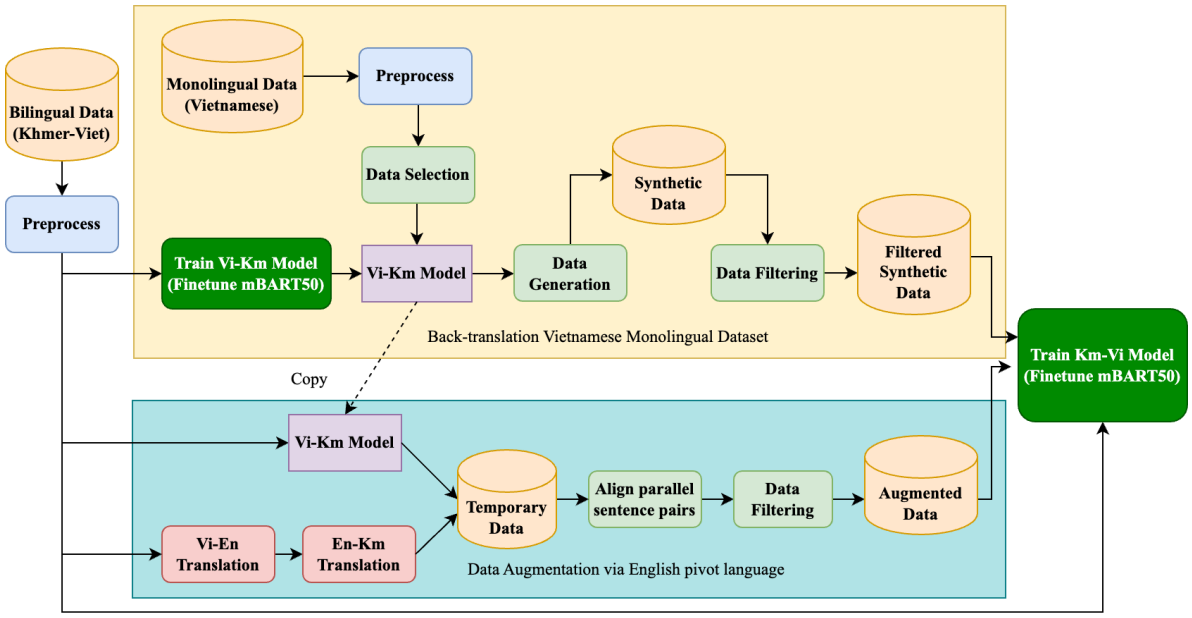

Figure 1: Our proposed Khmer-Vietnamese MT system diagram

semantic vector representations. This enables the identification of sentences in G that share the same domain as the sentences in D. Due to this reason, TF-IDF, cosine similarity, and their combination are utilized for ranking.

Data selection based on TF-IDF score

The term frequency (TF) measures the frequency of a term (word or subword) in a sentence, while inverse document frequency (IDF) is defined as the proportion of documents in the corpus that contain the term. So, TF-IDF score of a word w in a sentence s in G is calculated as:

$$score_w = TF - IDF_w = \frac{F_w^G}{W_s^G} \cdot \frac{T^D}{K_w^D}$$

where F_w^G is the frequency of w in s; W_s^G is the number of words in s; and K_w is the number of sentences in D contain w.

The TF-IDF score of the sentence $s \in G$ is evaluated as:

$$score_s^{(TF-IDF)} = \sum_{i=1}^{W_s^G} score_{w_i}$$

Data selection based on cosine similarity score The cosine similarity score between two sentences is calculated using a Bi-Encoder model [38]. This model includes a PLM combined with a pooler layer to encode each sentence as a sentence-level representation vector. Then, we compute the cosine similarity between these two vectors.

To choose the optimal PLM for the Vietnamese (target) language, we build a test set for the masked language model task, which includes 140,000 Vietnamese sentences from the Km-Vi bilingual dataset. Based on the accuracy of some well-known PLMs (ie, PhoBERT[1], XLM-RoBERTa[2], mDeBERTa[3]) using this dataset (Table 1), XLM-RoBERTa is selected as the PLM for the Bi-Encoder model.

Table 1: Accuracy of some models on the test set for the masked language model task.

Models	Accuracy
PhoBERT	80%
XLM-RoBERTa	87%
mDeBERTa	75%

The cosine similarity score of a sentence s in the G is calculated as:

$$score_s^{(COS)} = \frac{1}{|D|} \sum_{i=1}^{|D|} cos(s, D_i)$$

where $|D|$ is the number of sentences in D; D_i is the i-th sentence in D.

Data selection based on combination score

The combination score is calculated based on the TF-IDF score and the cosine similarity score:

$$score_s = \frac{score_s^{TF-IDF}}{\sum_{j=1}^{|G|} score_{G_j}^{TF-IDF}} + \frac{score_s^{COS}}{\sum_{j=1}^{|G|} score_{G_j}^{COS}}$$

where $|G|$ is the number of sentences in G; G_i is the i-th sentence in G

After assigning these scores to the sentences in the corpus G, the top 120,000 sentences from the target-language dataset with the highest score are selected to translate into the source language based on the target-to-source translation model.

[1]https://huggingface.co/vinai/phobert-base
[2]https://huggingface.co/xlm-roberta-base
[3]https://huggingface.co/microsoft/mdeberta-v3-base

4.1.2 Synthetic data generation

To increase the number of generated sentence pairs, each sentence from the target-language dataset is translated into k candidate sentences in the source language using the beam search (k is beam size) or top-k sampling method. As a result, k bilingual sentence pairs are created for each sentence in the target-language dataset. At this step, the synthetic dataset size can increase significantly. However, this dataset may contain many low-quality candidates. In the next section, we will propose our method to filter out the low-quality candidates.

4.1.3 Synthetic data filtering

Our data filtering approach is based on sentence-level cosine similarity. This approach involves comparing the similarity between the original sentence and its corresponding back-translated sentence, enabling us to identify and eliminate sentence pairs that exhibit significant deviations from the original meaning. Our method distinguishes itself from Koehn's approach [36] by not requiring an effective acquisition of the linear mapping relationship between the embedding spaces of the source and target languages. Instead, we leverage a cosine similarity measure to assess the semantic similarity between sentences.

Data filtering based on cosine similarity An important aspect of this approach is sentence representation in different languages. Although multilingual LMs (e.g., XLM-RoBERTa) are possible to do that, the representations for out-of-the-box sentences are rather bad. Moreover, the vector spaces of different languages are not aligned, meaning that words or sentences with the same meaning in different languages are represented in different vectors. Reimers and Gurevych [39] proposed a straightforward technique to ensure consistent vector spaces across different languages. This method uses a PLM as a fixed Teacher model that produces good representation vectors of sentences. The Student model is designed to imitate the Teacher model. It means the same sentence should be represented as the same vector in the Teacher model and the Student one. To enable the Student model to work with additional languages, it is trained on parallel (translated) sentences. The translation of each sentence should also be mapped to the same vector as the original one.

In Figure 2, the Student model should map "Hello World" and the German translation "Hallo Welt" to the vector of Teacher("Hello World"). This is achieved by training the Student model using the mean squared error (MSE) loss.

Based on this approach, we first generate two bilingual datasets: Vietnamese-English and English-Khmer parallel sentence pairs from the original Km-Vi dataset, using the Google Translator API. This API is taken from the deep translator [4]. The Student model is then trained on both the Vietnamese-English dataset and the Khmer-English one to create semantic vectors for three languages: English, Vietnamese, and Khmer. The representation vector of a sentence is the average of the token embeddings based on the Student model. We calculate the sentence-level cosine similarity between each parallel in the synthetic parallel dataset and filter out pairs with low scores.

[4]https://github.com/nidhaloff/deep-translator

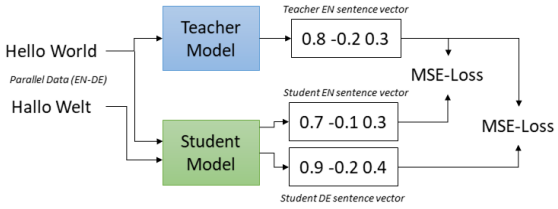

Figure 2: Given parallel data (e.g. English and German), train the student model such that the produced vectors for the English and German sentences are close to the teacher English sentence vector [39].

Data filtering using round-trip BLEU

The diagram of this method is represented in Figure 3. The process begins with the training of two NMT models: Km-Vi (source-to-target) and Vi-Km (target-to-source), using the given parallel sentence pairs. Next, we use the Vi-Km translation model to translate the monolingual sentences from the Vietnamese language to the Khmer one. We then take the translated sentences and back-translate them using the Km-Vi model. We evaluate the quality of sentence pairs based on sentence-level BLEU scores and discard sentence pairs with low scores.

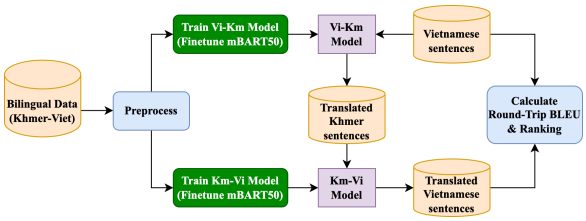

Figure 3: The diagram of the data filtering using round-trip BLEU.

4.2 Data augmentation method via english pivot language

A standard data augmentation method via English pivot language involves the translation of sentences in the target language from the original source-target parallel sentence pairs into English sentences. These English sentences are then translated into the source language to generate the source-target augmentation bilingual sentence pairs.

We propose an **"aligned"** version to improve the quality of the augmentation dataset. Given the original source-target sentence pair with a source sentence w_s and a target sentence w_t, we generate additional candidate sentences in the following way. The target sentence w_t is translated into

the source language using English pivot one. This step produces a candidate sentence in the source language w_{c1}. The target-to-source translation model described in Section 4.1 is used to generate another candidate sentence in the source language w_{c2}. The candidate pairs w_{c1} and w_{c2} are aggregated to get a temporary dataset. We carry out two filtering steps to remove low-quality parallel sentence pairs: (i) align parallel sentence pairs and (ii) data filtering. In the first step, the temporary dataset is aligned by three tools: Vecalign[5], Bleualign[6], and Hunalign [7]. Vecalign utilizes word embeddings to align sentences based on semantic similarity. Bleualign, on the other hand, uses the BLEU metric and n-gram overlap to align sentences in bilingual corpora. Hunalign is a heuristic-based tool that aligns parallel texts based on sentence length and lexical similarity. Sentence pairs that are aligned by two-third of the tools are selected to generate an aligned dataset. In the second step, the aligned dataset is filtered out based on the data filtering method in Section 4.1.3. As a result, we get an augmented dataset, which is combined with the synthetic parallel dataset from Section 4.1 and the original bilingual dataset to form the final training dataset.

5 Experiments

5.1 Experiment setup

We fine-tuned the mBART50 model on an RTX 3090 (24GB) GPU with different hyperparameters to choose the optimal parameter set for the model as follows: Adam optimization ($learning\_rate = 3\text{e}{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1\text{e}{-8}$) with linear learning rate decay scheduling. The best set of hyperparameters is employed in all our experiments.

To evaluate the effectiveness of our experiments, we used the BLEU score [40] through sacreBLEU[8] - an implementation version to compute the BLEU score. A higher BLEU score indicates better translation quality.

5.2 Experimental scenarios

To evaluate the effectiveness of our proposed methods for low-resource NMT, we used the Km-Vi biligual dataset from Nguyen et al. [16]. This dataset consists of 142,000 parallel sentence pairs, which were divided into a training set of 140,000 sentence pairs and a test set of 2,000 ones. In order to prevent biased phenomena in experiments, Nguyen et al. [16] randomly selected 2,000 sentence pairs from the original bilingual dataset to form the test set, following the distribution ratio of domains and lengths.

Six scenario groups were carried out in our experiments.

Scenario group #1 - Baseline model: Fine-tune the mBART50 model on the original Km-Vi bilingual dataset (Scenario #1).

All scenario groups from #2 to #6 used additional bilingual datasets which were generated from the Vietnamese corpus or the Km-Vi original bilingual one. This dataset was combined with the original dataset to create a larger training corpus. The Vietnamese dataset were created by crawling from online news websites (i.e, vnexpress.net[9], dantri.com.vn[10]), then preprocess to remove noise and long sentences. The langdetect[11] library was used to filter out non-Vietnamese sentences.

Scenario group #2 (#2.1 to #2.6) - Combine Scenario #1 and Back-translation: To generate a synthetic parallel dataset, 120,000 sentences from the above mentioned Vietnamese dataset were selected using our data selection strategies. These sentences were then translated into the Khmer language by using our back-translation method. We implemented and compared four data selection methods and two decoding ones (i.e, sampling and beam search).

Scenario group #3 (#3.1 to #3.3) - Combine Scenario #2 and Data filtering: In this scenario, we compared two methods in the data filtering strategy: the Round-Trip BLEU [35] (#3.1) and our proposed sentence-level cosine similarity (#3.2) . We experimented with two types of data selection: TF-IDF (#3.1 and #3.2) and combination score(#3.3).

Scenario group #4 (#4.1 to #4.2) - Combine Scenario #1 and Data augmentation via English pivot language: We compared "standard" and "aligned" versions to generate an augmented dataset. The Google Translator API is used for the translation task.

Scenario group #5 (#5.1 to #5.2) - Combine Scenarios #3 and #4: We created a new training dataset through the best settings from **Scenarios #3** and **#4**.

Scenario group #6 (#6.1 to #6.2)- Combine Scenario #5 and Data Generation: In this experiment, at the back-translation step, each sentence from the Vietnamese dataset was translated into k corresponding Khmer candidate sentences. Then these sentences were filtered and combined with the original bilingual dataset to create a new training dataset.

6 Experimental results

This section presents a comprehensive evaluation of our system performance under various scenarios and compares the best results with other relevant research. The analysis of the augmented data's quality is provided in Appendix 1.

6.1 Analysis our system performance using different scenarios

We evaluated our different scenarios on a test set with 2,000 parallel sentence pairs. The results of our scenarios are pre-

[5]https://github.com/thompsonb/vecalign
[6]https://github.com/rsennrich/Bleualign
[7]https://github.com/danielvarga/hunalign
[8]https://github.com/mjpost/sacrebleu
[9]https://vnexpress.net
[10]https://dantri.com.vn
[11]https://pypi.org/project/langdetect

Table 2: Experimental results

Scenario	Name	Data Augmentation Methods				BLEU (%)
		Back-translation			via English pivot language	
		Data Selection	Decoding Strategy	Data Filtering		
#1	Baseline model	-	-	-	-	52.32
#2.1	#1 + Back-translation	Randomness	Beam search	-	-	53.16
#2.2	#1 + Back-translation	Randomness	Sampling	-	-	53.49
#2.3	#1 + Back-translation	TF-IDF	Beam search	-	-	53.83
#2.4	#1 + Back-translation	TF-IDF	Sampling	-	-	53.96
#2.5	#1 + Back-translation	Cosine similarity	Sampling	-	-	53.98
#2.6	**#1 + Back-translation**	**Combination Score**	**Sampling**	-	-	**54.08**
#3.1	#2 + Data Filtering	TF-IDF	Sampling	Round-Trip BLEU	-	54.27
#3.2	#2 + Data Filtering	TF-IDF	Sampling	Cosine similarity	-	54.38
#3.3	**#2 + Data Filtering**	**Combination Score**	**Sampling**	**Cosine similarity**	-	**54.48**
#4.1	#1 + Data Augmentation	-	-	-	Standard	52.98
#4.2	**#1 + Data Augmentation**	-	-	-	**Aligned**	**53.29**
#5.1	#3 + #4	TF-IDF	Sampling	Cosine similarity	Standard	54.51
#5.2	**#3 + #4**	**Combination Score**	**Sampling**	**Cosine similarity**	**Aligned**	**54.93**
#6.1	#5 + Data Generation	Combination Score	Sampling	Cosine similarity	Standard	55.13
#6.2	**#5 + Data Generation**	**Combination Score**	**Sampling**	**Cosine similarity**	**Aligned**	**55.37**

sented in Table 2. The baseline **Scenario #1** achieved a 52.32% BLEU score.

Scenario group #2 shows that the combination score gave the best results and the sampling decoding method is better than the beam search method.

Table 3: Effect of BLEU filtering threshold in the data filtering using round-trip BLEU in the **scenario #3.**

Scenario	Threshold	BLEU (%)
#3	10	54.02
#3	**15**	**54.27**
#3	20	54.16
#3	25	53.80

For **scenario groups #3**, first, we evaluated the effect of data filtering thresholds to the system's performance. Tables 3 and 4 show that the BLEU score increases when the filter threshold is increased, but up to a certain threshold, and then reduced. This means that as the filter thresholds increase, we can filter out more low-quality parallel sentence pairs in the synthetic bilingual dataset, but the size of this dataset decreases. The best thresholds were then applied for all scenarios in groups #3 in order to compare the system performance with other scenarios in Table 2.

Scenario #4 First, in the standard version, we evaluated the model's performance with different augmented sizes. The original bilingual dataset was combined with 30000, 50000, and 70000 augmented sentence pairs created by the data augmentation via the English pivot language to form three training datasets. The obtained BLEU scores gradually increased from 52.48%, 52.52%, to 52.98%, proportional to the enhanced data size. The best result using 70000 augmented sentence pairs was used to compare with other scenarios in Table 2 (Scenario #4.1). Scenario #4.2 also used 70000 augmented sentence pairs in the aligned version.

Table 4: Effect of the cosine filtering threshold in the data filtering using sentence-level cosine similarity in the **scenario #3**.

Scenario	Threshold	BLEU (%)
#3	0.5	54.02
#3	0.6	54.36
#3	**0.7**	**54.38**
#3	0.8	53.92

With a result of 54.93% BLEU score, **Scenario #5** shown the effectiveness when combined the best synthetic parallel datasets from Scenario #3 and 30,000 pair sentences augmented in Scenario #4.

Finally, **Scenario #6**, we incorporated Scenario #5 with our generation strategy to get 55.37% BLEU points, which improved 3.05% BLEU scores compared to the baseline model. The results shown that the process of generating a synthetic dataset based on only one candidate with the highest probability was not enough. Taking k candidates and evaluating them helped us to retain more suitable candidates.

6.2 Comparison with other models

In addition to our scenario results above, we compared our best result with some models: Google Translator[12], pretrained multilingual seq2seq models, including mBART50 [37], m2m100-1.2B [41], and nllb-$*$ [42]-a multilingual translation model introduced by the Facebook AI[13] recently. The results shown in Table 5 indicated that our best model achieved best results for translating from the Khmer language to the Vietnamese one. In addition, our current

[12]https://github.com/nidhaloff/deep-translator

[13]https://ai.facebook.com/

approach had a better performance than our previous model [18] with 0.86% BLEU score higher.

Table 5: Comparison our system results to other models

Models	BLEU (%)
facebook/mbart50	12.74
facebook/m2m100-1.2B	22.44
facebook/nllb-200-distilled-600M	32.48
facebook/nllb-200-distilled-1.3B	36.51
facebook/nllb-200-3.3B	37.81
Google Translator	50.07
Our previous work [18]	54.51
Our best model	**55.37**

7 Conclusions

This research presents an approach to address the low-resource challenge in Khmer-Vietnamese NMT. The proposed method utilizes the pretrained multilingual model mBART as the foundation for the MT system, complemented by various data augmentation strategies to enhance system performance. These augmentation strategies encompass back-translation, data augmentation through an English pivot language, and synthetic data generation. The highest performance is achieved when combining the aforementioned augmentation methods with effective data selection and data filtering strategies, resulting in a significant 3.05% increase in BLUE score compared to the baseline model utilizing mBART with the original dataset. Our proposed approach outperforms the Google Translator model by 5.3% BLEU score on a test set of 2,000 Khmer-Vietnamese sentence pairs. Future work involves applying our proposed approach to other low-resource language pairs to demonstrate its generalizability.

References

[1] T. Khanna, J. N. Washington, and et al. Recent advances in apertium, a free/open-source rule-based machine translation platform for low-resource languages. *Machine Translation*, Dec 2021. https://doi.org/10.1007/s10590-021-09260-6.

[2] P. Koehn, F. J. Och, and et al. Statistical phrase-based translation. In *In Proceedings of NAACL*, page 48–54, 2003. https://doi.org/10.3115/1073445.1073462.

[3] P. Koehn, H. Hoang, and et al. Moses: Open source toolkit for statistical machine translation. pages 177–180. Association for Computational Linguistics, 2007. https://doi.org/10.3115/1557769.1557821.

[4] K. Cho, B. Merriënboer, and et al. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of EMNLP*, pages 103–111, 2014. https://doi.org/10.3115/v1/w14-4012.

[5] D. Suleiman, W. Etaiwi, and A. Awajan. Recurrent neural network techniques: Emphasis on use in neural machine translation. In *Informatica*, 2021. https://doi.org/10.31449/inf.v45i7.3743.

[6] Y. Tian, S. Khanna, and A. Pljonkin. Research on machine translation of deep neural network learning model based on ontology. In *Informatica*, 2021. https://doi.org/10.31449/inf.v45i5.3559.

[7] S. Edunov, M. Ott, and et al. Understanding back-translation at scale. In *Proceedings of EMNLP*, pages 489–500, 2018. https://doi.org/10.18653/v1/d18-1045.

[8] A. Vaswani, N. Shazeer, and et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. https://doi.org/10.48550/arXiv.1706.03762.

[9] Y. Chen, Y. Liu, and et al. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of ACL (Volume 1: Long Papers)*, pages 1925–1935, 2017. https://doi.org/10.18653/v1/p17-1176.

[10] Y. Kim, P. Petrov, and et al. Pivot-based transfer learning for neural machine translation between non-English languages. In *Proceedings of EMNLP-IJCNLP*, pages 866–876, 2019. https://doi.org/10.18653/v1/d19-1080.

[11] R. Sennrich, B. Haddow, and et al. Improving neural machine translation models with monolingual data. In *Proceedings of ACL (Volume 1: Long Papers)*, pages 86–96, 2016. https://doi.org/10.18653/v1/p16-1009.

[12] J. Zhang and C. Zong. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of EMNLP*, pages 1535–1545, 2016. https://doi.org/10.18653/v1/d16-1160.

[13] L. Mike, L. Yinhan, and et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, 2020. https://doi.org/10.18653/v1/2020.acl-main.703.

[14] C. Raffel, N Shazeer, A. Roberts, and et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. https://doi.org/10.48550/arXiv.1910.10683.

[15] Y. Liu, J. Gu, N. Goyal, and et al. Multilingual denoising pre-training for neural machine translation. *Transactions of ACL*, 8:726–742, 2020. https://doi.org/10.1162/tacl_a_00343.

[16] Van-Vinh Nguyen, , Huong Le-Thanh, and et al. KC4MT: A high-quality corpus for multilingual machine translation. In *Proceedings of LREC*, page 5494–5502, 2022.

[17] N. H. Quan, N. T. Dat, N. H. M. Cong, and et al. ViNMT: Neural machine translation toolkit, 2021. https://doi.org/10.48550/arXiv.2112.15272.

[18] V.H Pham and Le T.H. Improving khmer-vietnamese machine translation with data augmentation methods. In *Proceedings of SoICT '22*, pages 276–282, 2022. https://doi.org/10.1145/3568562.3568646.

[19] J. Devlin, M. Chang, and et al. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL: Human Language Technologies*, pages 4171–4186, 2019. http://doi.org/10.18653/v1/n19-1423.

[20] J. Zhu, Y. Xia, L Wu, and et al. Incorporating bert into neural machine translation, 2020. https://openreview.net/forum?id=Hyl7ygStwB.

[21] S. Rothe, S. Narayan, and et al. Leveraging pretrained checkpoints for sequence generation tasks. *Transactions of ACL*, 8:264–280, 2020. https://doi.org/10.1162/tacl_a_00313.

[22] B. Zoph, D. Yuret, and et al. Transfer learning for low-resource neural machine translation. In *Proceedings of EMNLP*, pages 1568–1575, 2016. https://doi.org/10.18653/v1/d16-1163.

[23] J. Hu, L. Zhang, and D. Yu. Improved neural machine translation with paraphrase-based synthetic data. In *Proceedings of NAACL*, 2019.

[24] X. Niu and et al. Subword-level word-interleaving data augmentation for neural machine translation. In *Proceedings of EMNLP*, 2018.

[25] Z. Liu and et al. Word deletion data augmentation for low-resource neural machine translation. In *Proceedings of ACL*, 2021.

[26] H. Wang and et al. Multi-objective data augmentation for low-resource neural machine translation. In *Proceedings of IJCAI*, 2019.

[27] C. Chu and et al. Domain adaptation for neural machine translation with limited resources. In *Proceedings of EMNLP*, 2020.

[28] M. Johnson, M. Schuster, and et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of ACL*, 5:339–351, 2017. https://doi.org/10.1162/tacl_a_00065.

[29] R. C. Moore and W. Lewis. Intelligent selection of language model training data. pages 220–224. Proceedings of ACL, 2010. https://aclanthology.org/P10-2041.

[30] M. Wees, A. Bisazza, and et al. Dynamic data selection for neural machine translation. In *Proceedings of EMNLP*, pages 1400–1410, 2017. https://doi.org/10.48550/arXiv.1708.00712.

[31] R. Wang, A. Finch, and et al. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of ACL*, pages 560–566, 2017. https://doi.org/10.18653/v1/p17-2089.

[32] S. Zhang and D. Xiong. Sentence weighting for neural machine translation domain adaptation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3181–3190, August 2018. https://aclanthology.org/C18-1269.

[33] C. C. Silva, C. Liu, and et al. Extracting in-domain training corpora for neural machine translation using data selection methods. In *Proceedings of the Third Conference on Machine Translation*, pages 224–231, 2018. https://doi.org/10.18653/v1/w18-6323.

[34] A. Poncelas and et al. Data selection with feature decay algorithms using an approximated target side. 2018. https://doi.org/10.48550/arXiv.1811.03039.

[35] A. Imankulova, T. Sato, and et al. Improving low-resource neural machine translation with filtered pseudo-parallel corpus. pages 70–78. Asian Federation of Natural Language Processing, 2017. https://aclanthology.org/W17-5704.

[36] P. Koehn, H. Khayrallah, and et al. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation*, pages 726–739, 2018. http://doi.org/10.18653/v1/w18-6453.

[37] Y. Tang, C. Tran, X. Li, and et al. Multilingual translation with extensible multilingual pretraining and finetuning, 2020. https://doi.org/10.48550/arXiv.2008.00401.

[38] J. Cho, E. Jung, and et al. Improving bi-encoder document ranking models with two rankers and multi-teacher distillation. In *Proceedings of SIGIR '21*, page 2192–2196, 2021. https://doi.org/10.1145/3404835.3463076.

[39] N. Reimers and I. Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation, 2020. https://doi.org/10.48550/arXiv.2004.09813.

[40] K. Papineni, S. Roukos, and et al. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318, 2002. http://doi.org/10.3115/1073083.1073135.

[41] A. Fan, S. Bhosale, H. Schwenk, and et al. Beyond english-centric multilingual machine translation, 2020. https://doi.org/10.48550/arXiv.2010.11125.

[42] NLLB Team. No language left behind: Scaling human-centered machine translation, 2022. https://doi.org/10.48550/arXiv.2207.04672.

A Appendix 1

To assess the quality of the augmented data, we present exemplary outputs from two methods described in Section 4 in Tables 6 and 7.

Table 6 exhibits examples generated by the back-translation method. Vi-Km sentence pairs in the first and second columns are added to the augmented training dataset if they pass the synthetic data filtering step. The table reveals that the NMT models employed in the back-translation process may still produce semantically incorrect sentences, particularly when translating proper names. Such sentences are subsequently filtered out during the data filtering process. Notably, no sentence pairs in the augmented dataset by this method exhibit poor quality.

In Table 7, we present examples of data augmentation via the English pivot language. Due to the relatively high quality of Google Translator, the augmented Vi-Km sentence pairs demonstrate a relatively high quality when the original Km-Vi sentence pair possesses good quality. However, discrepancies arise when the original Vietnamese-Khmer sentence pairs do not maintain complete semantic equivalence, leading to a similar outcome for the newly generated Khmer sentence and the original Vietnamese one. Consequently, in such instances, the data filtering step excludes the incorporation of the new sentence pair into the augmented dataset.

Table 6: Output examples of synthetic data generation process

Vi sentence (Vietnamese Monolingual Dataset)	**Km sentence generated by the Vi-Km model**	**Vi sentence generated by the Km-Vi model**	**Action**
Việt Nam đã thâm nhập và mở rộng thương mại tại thị trường này./*Vietnam has penetrated and expanded its trade in this market.*	វៀតណាម បាន ជ្រៀត ចូល និង ពង្រីក ពាណិជ្ជកម្ម ក្នុង ទីផ្សារ នេះ ។/*Vietnam has penetrated and expanded trade in this market.*	Việt Nam đã tham gia và mở rộng thương mại tại thị trường này./*Vietnam has joined and expanded trade in this market.*	Keep
Đoàn đại biểu kiều bào đã đến dâng hương ở tượng đài Vua Lê./*The overseas Vietnamese delegation came to offer incense at the statue of King Le.*	គណៈប្រតិភូ មកពី ប្រទេស វៀតណាម បាន មក បួងសួង នៅ រូបសំណាក ស្តេច Li Lei ។/*A delegation from Vietnam came to pray in the statue of King Li Lei.*	Một phái đoàn từ Việt Nam đã đến thăm các khu vực của Hoàng gia Li Lei./*A delegation from Vietnam visited the areas of Royal Li Lei.*	Filter out
Theo đó, các dụng cụ này dao động mức từ vài chục cho đến hàng chục triệu đồng./*Accordingly, these tools range from a few tens to tens of millions of dong.*	តាមរយៈ នេះ ឧបករណ៍ ទាំងនេះ មាន តម្លៃ ពី មួយ ដង ទៅ មួយ ដង ទៅ មួយ ដង ។/*Through this, these devices are priced from time to time.*	Bằng cách này, những thiết bị này có giá trị một lần, một lần, một lần./*This way, these devices are worth it once, once, once.*	Filter out

Table 7: Output examples of data augmentation process via english pivot language

Original Km sentence	**Original Vi sentence**	**Km augmented sentence**	**Action**
កញ្ចប់ ទិន្នន័យ ត្រូវបាន តម្រៀប តាម តំបន់ សម្រាប់ អ្នកទិញ ងាយស្រួល ជ្រើសរើស ។/*The data packets are sorted by area for the buyer to easily select.*	Các gói data được chia ra theo khu vực để người mua dễ dàng lựa chọn./*The data packages are divided by region for buyers to easily choose*	កញ្ចប់ ទិន្នន័យ ត្រូវបាន ចាត់ ថ្នាក់ តាម តំបន់ ដើម្បី ងាយស្រួល ជ្រើសរើស អ្នកទិញ ។/*Data packages are categorized by area for easy selection of buyers.*	Keep
មនុស្ស ប្រមាណ ៥០០ លាន នាក់ អាច ប្រឈម នឹង ភាពក្រីក្រ ដោយសារ វិបត្តិ សេដ្ឋកិច្ច ដ៏ ធ្ងន់ធ្ងរ បំផុត តាំងពី មុន មក ។/*An estimated 500 million people could face poverty due to the worst economic crisis ever.*	Thế giới đang đối mặt với cuộc suy thoái kinh tế sâu sắc nhất, được đánh giá là nghiêm trọng hơn các cuộc khủng hoảng trước đây./*The world is facing the deepest economic recession, which is considered to be more severe than previous crises.*	ពិភពលោក កំពុង ប្រឈមមុខ នឹង វិបត្តិ សេដ្ឋកិច្ច ដ៏ ជ្រៅ បំផុត ដែល ត្រូវបាន គេ ចាត់ទុក ថា ធ្ងន់ធ្ងរ ជាង វិបត្តិ មុន ។/*The world is facing the deepest economic recession, which is considered to be more severe than previous crises.*	Filter out