Journal of Computer Science and Cybernetics, V.41, N.3 (2025), 285-304
DOI: 10.15625/1813-9663/21055

VIBIDIRECTIONMT-EVAL: MACHINE TRANSLATION FOR VIETNAMESE - CHINESE AND VIETNAMESE - LAO LANGUAGE PAIR

TRAN HONG VIET, NGUYEN MINH QUY, NGUYEN VAN VINH*

University of Engineering and Technology
Vietnam National University, Hanoi, Vietnam
144 Xuan Thuy Street, Cau Giay Ward, Ha Noi, Viet Nam

Abstract. This paper presents the results of the VLSP 2022-2023 Machine Translation Shared Tasks, focusing on Vietnamese-Chinese and Vietnamese-Lao machine translation. The tasks were organized as part of the 9th and 10th annual workshops on Vietnamese Language and Speech Processing (VLSP 2022, VLSP 2023). The objective of the shared task was to build machine translation systems, specifically targeting Vietnamese-Chinese and Vietnamese-Lao translation (corresponding to 4 translation directions). The submissions were evaluated on 1,000 test pairs from both news and general domains, using established metrics such as BLEU and SacreBLEU. In addition to these automated evaluations, the system outputs were also assessed through human judgment by experts in the Chinese and Lao languages. These human assessments played a crucial role in ranking the performance of the machine translation models, ensuring a more comprehensive evaluation.

Keywords. Machine translation, neural machine translation, low-resource, translation error analysis.

1. INTRODUCTION

Neural Machine Translation (NMT) has currently obtained state-of-the-art in machine translation systems. However, the translation quality is still a challenge in translation systems. Neural Machine Translation (NMT) [1, 2, 3] has recently shown impressive results compared to Statistical Machine Translation (SMT) [4, 5]. However, NMT systems still have great challenges [6]. The MT track basically corresponds to a subtitling translation task. The natural translation unit considered by the human translators volunteering for News is indeed the single caption - as defined by the original transcript - which in general does not correspond to a sentence, but to fragments of it that fit the caption space. While translators can look at the context of the single captions, arranging the MT task in this way would make it particularly difficult, especially when word reordering across consecutive captions occurs. For this reason, we preprocessed all the parallel texts to rebuild the original sentences, thus simplifying the MT task. Table 1 provides statistics on in-domain texts supplied for training and evaluation purposes for each MT task. Texts are pre-processed (to-

*Corresponding author.
E-mail addresses: thviet@vnu.edu.vn (T.H. Viet); minhquy1624@gmail.com (N.M. Quy); vinhnv@vnu.edu.vn (N.V. Vinh).

kenization, Chinese and Vietnamese segmentation) with the tools used for setting-up baseline systems (see below). For this purpose, the task involved creating a comprehensive dataset with human-annotated translations[1].

In 2022, a significant milestone was reached for Machine Translation (MT) within the VLSP evaluation campaign, particularly in Chinese - Vietnamese translation through news sources. Although data scarcity posed a major challenge, participating teams successfully developed specialized methods by leveraging the linguistic similarities between Chinese and Vietnamese. Notably, the one-to-one mapping between Sino-Vietnamese and Chinese words played a crucial role in their success. Moving into 2023, VLSP has shifted its focus to Lao - Vietnamese and Vietnamese - Lao Machine Translation tasks, where limited training data remains a significant barrier. However, the close linguistic relationship between Lao and Vietnamese, including numerous one-to-one lexical correspondences, presents opportunities to apply similar specialized techniques. These approaches hold promise even in scenarios with constrained data availability.

2. TRAINING AND TEST DATA

In the VLSP 2022 and VLSP 2023 evaluation campaigns, we released comprehensive training datasets designed to support Vietnamese-Chinese and Vietnamese-Lao machine translation tasks. These datasets include development and public test sets to facilitate model optimization and evaluation. Specifically, the VLSP 2022 dataset comprises over 300,000 Vietnamese-Chinese bilingual sentence pairs for training, with an additional 1,000 sentences for development and testing. Similarly, the VLSP 2023 dataset, designed for Vietnamese-Lao translation, contains 100,000 bilingual sentence pairs for training, 2,000 for development, and 1,000 for testing. The provision of development and public test sets allows

Table 1: Bilingual training and evaluation corpora statistics

Task	Dataset	Sent	Tokens [7]		
			vi	zh	lo
Vi ↔ Zh	Train	300,348	43,762	141,879	-
	Dev	1,000	2,545	3,796	-
	Test	1,000	2,454	4,078	-
Vi ↔ Lo	Train	100,000	227,000	-	120,710
	Dev	2,000	4,262	-	1,740
	Test	1,000	2,454	-	3,126

participants to fine-tune their models before formal evaluation on the secure private test set. Notably, all development, public test, and private test sets are within the same linguistic domain, ensuring consistency in evaluation and benchmarking. SacreBLEU is recommended for model evaluation, as it offers a reliable metric for assessing machine translation accuracy.

The input data is provided in UTF-8 text format, with 1-to-1 aligned bilingual sentence pairs, ensuring precise correspondence throughout training and testing. This approach facil-

[1]https://huggingface.co/datasets/VLSP2023-MT/ViBidirectionMT-Eval

itates standardization and improves the accuracy of machine translation systems, contributing to research and application in automatic translation for Vietnamese in a multilingual context.

Table 1 shows statistics on in-domain texts supplied for training and evaluation purposes for two MT tasks: Vietnamese ↔ Chinese Machine Translation Systems for VLSP 2022 and Vietnamese ↔ Lao Machine Translation Systems for VLSP 2023. All parallel texts were tokenized and truncated using sentence piece scripts, and then they are applied to Sennrich's BPE [7]. For Vietnamese, we only apply Moses's scripts for tokenization and true-casing.

3. EVALUATION

The participants to the MT track had to provide the automatic translation of the test sets in text format. The output had to be case-sensitive, detokenized and had to contain punctuation. The quality of the translations was measured both automatically, against the human translations created by the open translation project, and via human evaluation (Section 5).

Case sensitive scores were calculated for the three automatic standard metrics BLEU [8] and SacreBLEU [9], as implemented in mteval-v13a.pl and sacrebleu, by calling:

- mteval-v13a.pl -c
- sacrebleu -t vlsp2022/systems -l zh-vi –echo MTTracks
- sacrebleu -t vlsp2023/systems -l lo-vi –echo MTTracks

Detokenized texts were passed, since the two scorers apply an internal tokenizer. Before the evaluation, Chinese texts were segmented at character level, keeping non-Chinese strings as they are. In order to allow participants to evaluate their progress automatically and in identical conditions, an evaluation server was developed. Participants could submit the translation of any development set to either a REST Web service or through a GUI on the web, receiving as output the three scores BLEU, NIST [10], TER [11] and SacreBLEU computed as above. The core of the evaluation server is a shell script wrapping the mteval scorers. The evaluation server was utilized by the organizers for the automatic evaluation of the official submissions. After the evaluation period, the evaluation on test sets was enabled to all participants as well.

4. SYSTEM SUBMISSIONS

In the multilingual machine translation tasks at VLSP 2022 and VLSP 2023, we conducted Vietnamese-to-Chinese and Vietnamese-to-Lao translation tasks, attracting substantial participation from both domestic and international organizations. Specifically:

VLSP 2022 - MT: The machine translation task for Vietnamese-to-Chinese and vice versa had 25 registered teams, including universities such as JAIST and HUST, as well as major corporations like Samsung SDS, VinBigData, and VCCorp. Among these, 5 teams submitted official entries, complete with models for performance evaluation and detailed technical reports.

VLSP 2023 - MT: The Vietnamese to Lao and Lao to Vietnamese translation task attracted 26 registered teams, including institutions and universities such as HUST, MTA,

UET-VNU, and technology companies like Viettel and Bosch Global Software Technologies Vietnam. In total, 7 teams submitted official entries for evaluation.

In both machine translation tasks, we selected for each task the three methods that achieved the highest results for each translation task. Each of these methods has technical reports that demonstrate the approach, method content, contribution to the machine translation task, and results achieved. We present each of these methods in each translation task the following section in 4.1 and 4.2.

4.1. Vietnamese-Chinese machine translation

In the task of Vietnamese-Chinese bidirectional machine translation, we selected the three most effective approaches to achieve accurate and fluent translations that preserve the original meaning. Each method was carefully evaluated for its translation accuracy and fluency to ensure high-quality, natural output. The teams employed distinct techniques and strategies, including language model fine-tuning and input data optimization, to maximize the quality and naturalness of the translations. The selected methods are as follows:

- Team 1 (SDS): An efficient approach for machine translation on low-resource languages.
- Team 2 (VBD-MT): VBD-MT Vietnamese-Chinese bidirectional translation system.
- Team 3 (JNLP): An effective method using Phrase mechanism in Neural machine translation.

The advantage of the SDS and VBD-MT teams is that they both utilize the pre-trained mBART model (which is the noise encoder-decoder architecture, quite suitable for the machine translation task). Team SDS uses mBART-50 with input preprocessing, whereas MTA uses mBART-25 and applies post-processing to numerical and date data. However, the weakness for SDS and VBD-MT stems from the drawbacks of mBART itself: namely the need for significant GPU computational resources, the large number of parameters (mBART-50 has approximately 680 million parameters), and the difficulty in debugging translation errors compared to a vanilla Transformer (for a machine translation task, they could completely develop a vanilla Transformer model from scratch and optimize it specifically for this purpose). Additionally, the ensemble technique employed by VBD-MT is also quite time-consuming.

The advantage of the JNLP team is the integration of linguistic information (phrases) into the Transformer model. This could improve the translation quality of phrases and idioms. However, the weakness is that effective integration into the Transformer model still requires further experimentation, and the resulting impact on speed is also a factor to consider.

4.1.1. An efficient approach for machine translation on low-resource languages

The team proposes leveraging data synthesis as a technique to augment the training set for low-resource language pairs, particularly Vietnamese-Chinese. To accomplish this, the mBART-50 [12] machine translation system is first fine-tuned with existing bilingual data. It is then employed to translate from the target language back into the source language, effectively generating a synthetic bilingual dataset. This newly synthesized dataset is subsequently merged with authentic bilingual data, providing a more comprehensive training set for the final model.

In constructing the final translation model, the team follows a systematic approach involving three key steps: (1) Training a Vietnamese-English translation model with mBART-50; (2) Enhancing the dataset by generating additional bilingual data through selected sentences

extracted from the monolingual dataset; (3) Fine-tuning the model using this expanded bilingual dataset. The VLSP 2022 dataset, which includes 300,000 bilingual sentence pairs and extensive, cleaned monolingual corpora, is employed to ensure that the input data remains of high quality throughout the training process.

For low-resource language pairs, the team applies the TF-IDF selection technique to identify and extract significant sentences from a large monolingual dataset containing 25 million Vietnamese and 19 million Chinese sentences. The resulting dataset, a synthesized bilingual corpus, is then combined with the original bilingual data to enhance the accuracy and robustness of the final translation model.

The utilization of mBART-50 in this study capitalizes on its multilingual translation capabilities, achieved through denoising training. By supporting up to 50 languages and subsequently fine-tuning the model with VLSP data, the research team successfully developed high-quality multilingual machine translation models specifically tailored to Vietnamese and Chinese, thereby enhancing translation performance for these low-resource languages.

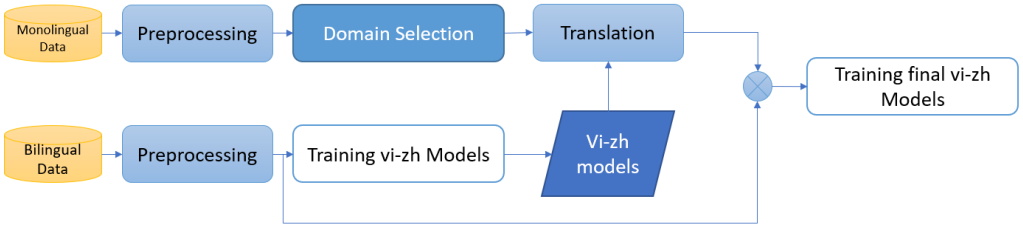

Figure 1: Flow of data processing and model training

The system depicted in the Figure 1 illustrates the training process of a machine translation model for low-resource language pairs, particularly Vietnamese and Chinese. First, both bilingual and monolingual data are processed, and key sentences are selected using the TF-IDF method to identify domain-specific content and ensure high-quality inputs. Subsequently, the mBART-50 model is fine-tuned on the existing bilingual dataset and used for back-translation from the target language to the source language, thereby creating a synthetic bilingual dataset. The authentic bilingual dataset is then combined with this synthetic dataset, forming a robust training corpus. This approach ultimately enhances the accuracy and stability of the translation system for Vietnamese and Chinese.

4.1.2. VBD-MT Vietnamese-Chinese bidirectional translation system

Baseline system is conducted using the robust Transformer model, which is fine-tuned with mBART-25, a model pre-trained on 25 languages, including both Chinese and Vietnamese. For text processing, we implement the SentencePiece tool to handle tokenization and vocabulary filtering, reducing the vocabulary size from an initial 250K to 67K tokens. This reduction aligns with the limited GPU resources available, allowing for efficient training without the need for high-performance server infrastructure.

To further enhance the dataset, the team apply back-translation using a top-k sampling technique, selecting the top 5 highest-scoring outputs to generate diverse synthetic data. This method yields 211K back-translated sentence pairs from Chinese to Vietnamese (Zh-Vi) and 403K pairs from Vietnamese to Chinese (Vi-Zh). The synthetic data is then combined with authentic bilingual data, effectively expanding the training set and boosting model performance.

The system also employs an ensemble method, achieved by averaging the model weights from the last N checkpoints, where N is optimally set to 5. This ensembling approach significantly enhances accuracy, particularly when used alongside the back-translation data.

To address potential translation errors in numeric and date-time values, the team introduce a post-processing step with customized patterns designed for these data types. Although this post-processing does not directly increase the BLEU score, it improves translation quality by ensuring that critical values, such as those related to people and currency, are accurately translated.

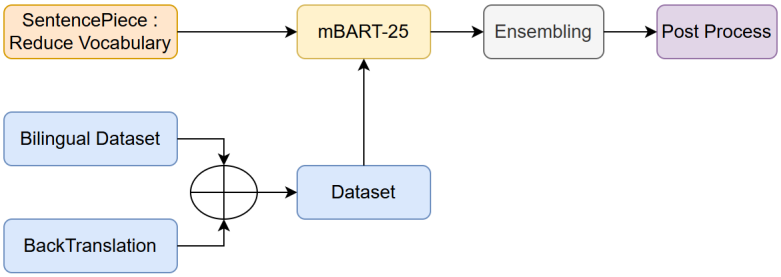

Figure 2: System flow machine translation

The system 2 fine-tunes a Transformer model in conjunction with mBART-25, where mBART-25 is utilized as a pre-trained multilingual model. SentencePiece is employed to reduce the vocabulary size, optimizing GPU resource utilization. Back-translation is applied to generate additional bilingual data by translating sentences from the target language back to the source language, which are then merged with the original dataset to expand the training corpus. An ensemble technique is used to enhance accuracy. Finally, a rule-based post-processing step corrects errors related to numerical data and dates, improving the system's output quality.

After evaluating various models, the team selected Fairseq [13] for the baseline system, as it demonstrated superior performance on the public test set. For Chinese-Vietnamese translation, the model achieved a BLEU score of 38.0, which improved to 38.8 with the inclusion of back-translation; for Vietnamese-Chinese translation, the BLEU score increased from 37.8 to 38.0 with the addition of both back-translation and ensembling.

Final system submission for the shared task integrates baseline modeling, back-translation, ensembling, and post-processing. The post-processing step, focused on automatically correcting numeric and date-time values, ensures a more accurate and higher-quality translation output for critical data, providing a well-rounded, effective solution.

4.1.3. An effective method using phrase mechanism in neural machine translation

This approach developed PhraseTransformer, a model based on the Transformer architecture that incorporates phrase-based attention mechanisms to improve machine translation performance. Unlike prior models, PhraseTransformer eliminates the need for external syntactic tree information, making it more efficient and lightweight compared to other phrase-level attention models. The core concept behind PhraseTransformer is to enhance word representations by leveraging local context and capturing dependencies between phrases within a sentence, enabling more nuanced translation outputs.

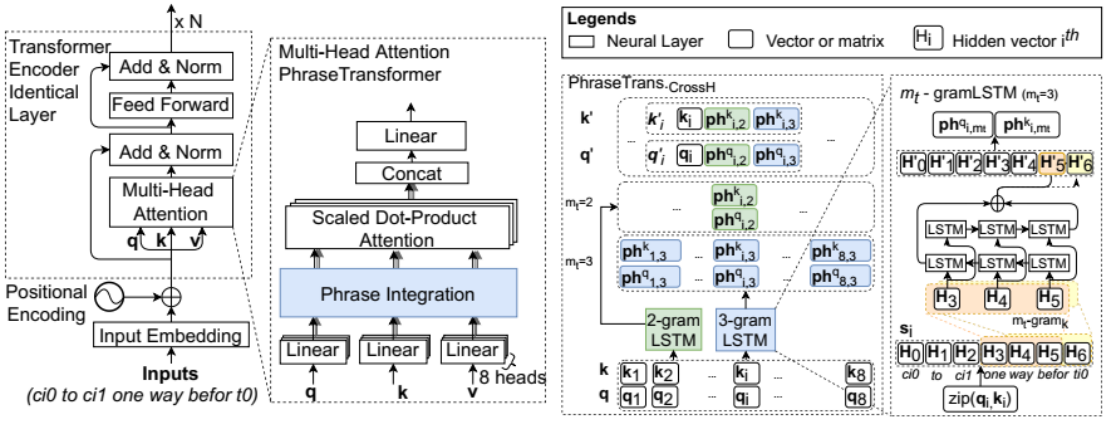

Figure 3: Overview of PhraseTransformer (CrossH) using n-gram LSTM in MultiHead layer. In this case, the phrase representations are built with gram_size = {2, 3}, 2-gram, 3-gram models apply to all 8 heads.

In the preprocessing stage, they utilized Byte-Pair Encoding (BPE) to address out-of-vocabulary issues by breaking words down into sub-tokens. For Vietnamese, this work performed 4,000 BPE operations, while for Chinese, which lacks inherent word spacing, they applied 16,000 operations. For Chinese text, the BPE segmentation module treats the entire raw sentence as a single word segment, ensuring effective sub-tokenization even without spacing between characters.

To evaluate PhraseTransformer's performance against the original Transformer model, they trained both models on the Chinese-Vietnamese bilingual dataset provided by VLSP 2022, without any supplementary external data or pretrained models. Both models were tested under identical configurations, and they averaged the weights from the last 5 checkpoints to produce the final model used for translation testing.

The experimental results reveal that PhraseTransformer consistently outperforms the original Transformer across various n-gram sizes, underscoring the effectiveness of its phrase-based attention mechanism in capturing sentence meaning. Furthermore, PhraseTransformer's adaptability extends beyond translation tasks to other languages and NLP applications, as it operates independently of external syntactic tree information, making it a versatile tool for diverse linguistic challenges.

The PhraseTransformer (CrossH) system depicted in the Figure 3 is a variant of the Transformer architecture that incorporates a phrase-based attention mechanism to enhance machine translation quality. This model employs n-gram LSTM within multi-head attention to capture local context and inter-phrase relationships, thereby obviating the need for external syntactic tree information.

4.2. Vietnamese-Lao machine translation

In the Vietnamese-Lao bidirectional machine translation task, high performance has been achieved using methods based on the Transformer architecture, a leading approach in machine learning. These methods exploit the Transformer's capacity to deliver high-quality translations by efficiently managing complex word sequences and capturing semantic relationships between words within sentences. Notably, by fine-tuning Transformer-based pretrained

models, we can tailor the system to better handle Vietnamese-Lao bilingual data, thereby improving translation accuracy and naturalness. This fine-tuning enhances the system's precision while also increasing its capacity to capture contextual meaning and accurately reproduce the unique grammatical structures of both languages, achieving high standards in bidirectional machine translation quality.

In the task of Vietnamese-Lao bidirectional machine translation, we selected the three most effective approaches, specifically as follows:

- Team 1 (BlueSky): A Transformer-based model for Lao-Vietnamese machine translation.
- Team 2 (MTA_AI): Vietnamese-Lao bidirectional translation system.
- Team 3 (BGSV AI): A sequence-to-sequence model for Lao-Vietnamese machine translation.

The advantage of the BlueSky team is utilizing the power of mBART. However, there are two weaknesses in their method that require careful adaptation without causing catastrophic forgetting of mBART's ability, and training custom mBART requires significant resources and high-quality mono data.

With the MTA AI team, the advantage of the team is that using powerful pre-trained multilingual models (m2m100, mT5) and large-scale back-translation (1.5M sentences) boosts performance. The weaknesses are that it heavily depends on Google Translate for synthetic data, which can have risks in quality inconsistency, and mT5-small may limit capacity compared to larger models.

The advantage of the BGSV AI team is building a new tokenizer from scratch and a detailed and efficient preprocessing method for Vietnamese text. This could improve the translation quality of phrases and idioms. However, the weakness is that it only employs basic preprocessing techniques on Lao text and does not utilize the benefits of using pretrained models.

4.2.1. A Transformer-based model for Lao-Vietnamese machine translation

Blue_Sky leverages a pretrained mBART model [12] initially trained on extensive monolingual datasets in both Vietnamese and Lao. The model is subsequently fine-tuned using a bilingual dataset from VLSP, which enhances its translation accuracy and fluency for both languages. This approach integrates diverse linguistic data from monolingual sources, allowing the model to capture complex grammatical and syntactical structures unique to Vietnamese and Lao, providing a strong base for the fine-tuning phase.

For machine translation tasks, the Transformer WMT [14] en-de big model was employed. This model utilizes an Encoder-Decoder architecture, where the Encoder processes the source sentence to gather context, and the Decoder generates the target sentence sequentially, one word at a time. The model leverages the Transformer's powerful self-attention mechanism to optimize translation accuracy while maintaining semantic consistency.

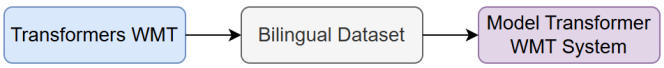

Figure 4: Training phases of mBART and Transformer WMT models

To further assess the capabilities of the large language model, this approach adapted mBART for Lao, as the original version of mBART does not support this language. The adaptation involved training mBART on monolingual Vietnamese and Lao datasets to extend its language support, followed by fine-tuning on the bilingual dataset provided by VLSP. This adaptation ensures the model's effectiveness in machine translation between Vietnamese and Lao.

This method also employed SentencePiece [15] for tokenizing Vietnamese and Lao text, setting a vocabulary size of 20,000 tokens. The training dataset consisted of 100,000 sentence pairs, with a test set of 2,000 pairs from VLSP used for evaluation. Additionally, the model was pretrained on a large monolingual dataset 1.8 GB of Vietnamese text and 1 GB of Lao text - laying a strong foundation for fine-tuning. However, the fine-tuned mBART model performed below the Transformer WMT en-de big model in terms of overall translation accuracy. System 4 utilizes the Transformer WMT, a conventional encoder-decoder architecture that leverages the self-attention mechanism to optimize machine translation on bilingual datasets.

4.2.2. Vietnamese-Lao bidirectional translation system

The MTA_AI team utilized the M2M-100 418M [16] and mt5-small [17] models to fine-tune a translation system for Vietnamese and Lao, both of which are low-resource languages with limited pre-trained model support. After an extensive survey of available multilingual models, they determined that m2m_100 [16] and mT5 [18, 17] were particularly well-suited for this project. These models are capable of translating multiple language pairs, including Vietnamese and Lao, making them ideal choices for enhancing translation quality between these languages.

The M2M100-418M model is a multilingual encoder-decoder designed for many-to-many translation, supporting direct translation between numerous languages without needing a pivot language. The mT5-small model, a compact version of T5 with a multilingual capability, was pre-trained on the Common Crawl dataset, covering 101 languages and comprising 300 million parameters. The model is fine-tuned using the Adam optimizer [19]. This combination allows both comprehensive language support and computational efficiency.

In this approach, MTA_AI team applied back-translation using Google Translate to convert monolingual sentences into bilingual data, thereby creating a synthetic dataset. This method generated 1.5 million sentence pairs for both Vietnamese-to-Lao (Vi-Lo) and Lao-to-Vietnamese (Lo-Vi) translations, substantially expanding our training data.

To investigate the impact of large-scale data on model performance, they trained the m2m100-418M model with a total of 3 million back-translated monolingual sentences. The results demonstrated a significant enhancement in translation accuracy, affirming the positive influence of large-scale data on the effectiveness of machine translation systems for low-resource languages.

The Figure 5 depicted in the figure illustrates how the team employs the `m2m_100-418M` and `mT5_small` models to train machine translation between Vietnamese and Lao using the back-translation method. They generate synthetic bilingual data from monolingual sentences, thereby expanding the training dataset. This approach significantly enhances the translation quality for these two low-resource languages.

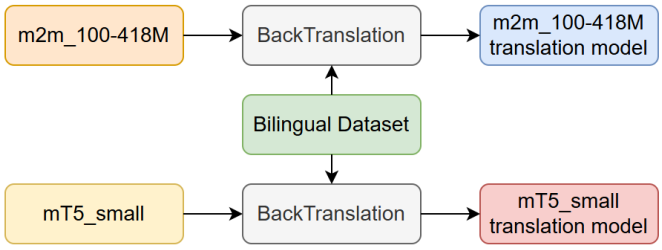

Figure 5: Training phases of mT5_small and m2m_100-418M models

4.2.3. A sequence-to-sequence model for Lao-Vietnamese machine translation

The BGSV_AI team employed a sequence-to-sequence approach, utilizing large language models to tackle the machine translation challenge in the shared-task competition. During the evaluation phase, they observed that existing models did not support both Vietnamese and Lao simultaneously. Consequently, they developed a unique tokenizer using the SentencePiece technique [15] to generate a tailored vocabulary set suited to both languages. They then customized and trained the T5 model [18] from scratch, specifically for this machine translation task.

Pre-processing proved essential in enhancing both translation quality and efficiency. This stage involved cleaning the dataset to remove noise, standardizing formats (such as dates and numbers) for uniformity, and tokenizer the text into smaller units. Given their limited familiarity with the Lao language, they applied only fundamental pre-processing techniques, which included the removal of irrelevant characters and symbols.

In the experiment, this approach focused on optimizing the tokenizer to improve sentence comprehension while managing the vocabulary size effectively. To achieve this, BGSV_AI team sets a token length of 90 for Lao and 150 for Vietnamese, aiming for an optimal balance between computational efficiency and language understanding. These customized token lengths were carefully tailored to the linguistic characteristics of each language, thereby maximizing the performance of their machine translation system.

The system illustrated in the Figure 6 represents a customized T5 model developed by the team for bidirectional translation between Lao and Vietnamese. They created separate tokenizers for both languages using SentencePiece, while also performing preprocessing and adjusting token lengths as appropriate. As a result, the model is trained from scratch to optimize the machine translation for both Lao and Vietnamese.

5. EXPERIMENTAL RESULTS

For all language pairs, we show the case-sensitive BLEU and SacreBLEU scores. Results would be ranked by human evaluation.

- Only constrained systems will be evaluated and ranked.
- Unconstrained systems would not be human-evaluated and ranked.
- By Human: expert in Chinese and Lao languages (05 expert).

VLSP 2022-MT We observed that all participant systems outperformed the baselines, with tasks involving Chinese and Vietnamese attracting particular attention. For Chinese

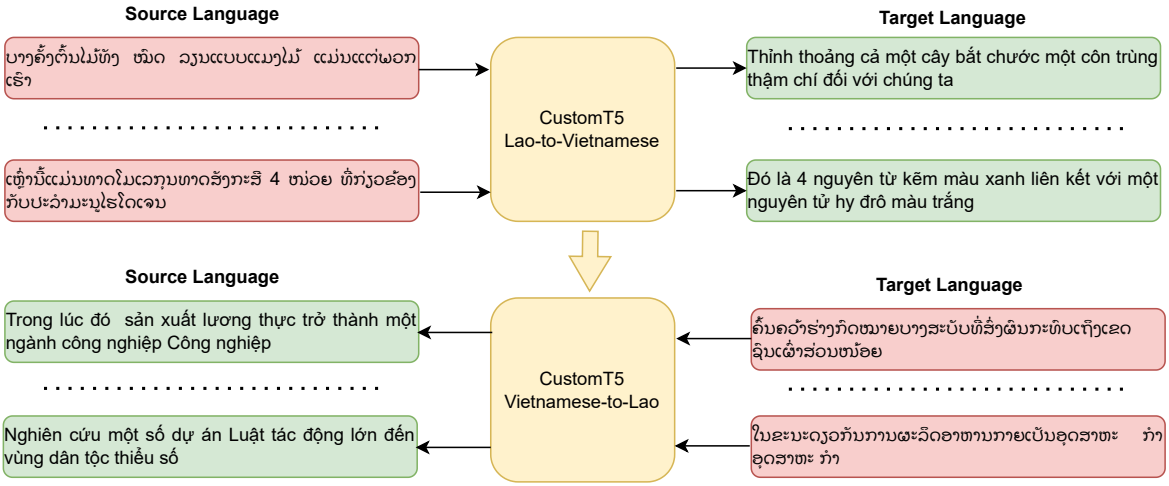

Figure 6: Overview of their proposed machine translation framework, which includes Lao-to-Vietnamese and Vietnamese-to-Lao directions.

Table 2: The summary of methodology for MT shared-task Chinese-Vietnamese

No	Teams	Methodology
1	VBD-MT	Pre-trained model mBART, using sampling method for backtranslation, applying ensembling and postprocessing to improve the translation quality.
2	SDS	Pre-trained language model mBART for the machine translation task, proposing data selection and data synthesis techniques from the monolingual corpus.
3	S-NLP	No technical report
4	VC-datamining	Using transformer model and integrating: data filtering, checkpoint averaging, data augmentation, ensemble.
5	JNLP	Using Phrase Transformer for incorporating the phrase dependencies information into the Self-Attention mechanism.

language, which is a language notoriously difficult to process, the better systems largely beat the basic methods featured in the baselines. For Vietnamese language, participant scores vary a lot as well; differently than on Chinese, submitted runs hardly provided higher quality than baselines; in particular, on Vietnamese-to-Chinese direction, none was able to improve the baseline translation: despite a deep analysis, we were unable to find a plausible explanation for this surprising outcome.

The Table 2 shows the methodology for the teams in VLSP 2022 MT shared-task Chinese-Vietnamese. In the Vietnamese-Chinese translation task, the VBD-MT team applied the pretrained models mBART25 [20] and mBART50 [12], with experimental results on the test set indicating that mBART25 achieves higher performance. The performance of the mBART

model has been presented in the study by Tang et al. (2020) [12]. In the report of the SDS team, the mBART50 model is used as the pretrained model for training on the bilingual dataset. The comparison of translation performance between the two teams is shown in Table 3, with the results indicating that the SDS team achieved better performance in this task.

Table 3: Vietnamese to Chinese machine translation task (Automatic evaluation results)

No	Team Name	BLEU	SacreBLEU
1	S-NLP	26.62	26.65
2	SDS	21.87	21.85
3	JNLP	21.70	21.76
4	VBD-MT	17.95	18.02
5	VC-datamining	17.10	17.15

In Table 3, the Chinese-Vietnamese translation direction, the top three teams significantly outperformed the remaining two teams. Although the S-NLP team, ranked third, scored much lower in BLEU compared to the top two teams, they achieved a substantially higher score in ScareBLEU.

For the Vietnamese-Chinese translation direction, the S-NLP team performed exceptionally well, surpassing the second-ranked team by nearly 5 points in both BLEU and ScareBLEU scores.

VLSP 2023-MT

The teams engaged in the Vietnamese-Lao machine translation task with the aim of enhancing the quality of translation models for this language pair. The methods implemented by the teams significantly outperformed baseline models on the test dataset, especially under human evaluation. The summary of methodology machine translation in VLSP 2023 for Lao-Vietnamese language pair given by in the Table 4.

In Table 5 with the Lao-Vietnamese translation direction, three teams achieved significantly higher scores compared to the others. Notably, the human evaluation scores for these teams were exceptionally high, playing a decisive role in determining their rankings.

In Table 6 the Viet-Lao translation direction, two teams achieved outstanding results compared to the others, both scoring more than 50 points. Notably, the MTA_AI team achieved a score of 61. In the final results, the human evaluation scores were decisive in determining the rankings. Although the MTA_AI team had a lower ScareBLEU score, their human evaluation score was exceptional.

6. HUMAN EVALUATION

To accurately and comprehensively evaluate the quality of the translation model, we decided to leverage the expertise and experience of specialists in the field of linguistics. The evaluation process began by obtaining sentence translations from various models and then submitting these translations to experts for review and scoring. This process went beyond merely comparing results with standard translations; it required experts to analyze and assess based on multiple factors such as semantic accuracy, syntax, context, and naturalness of the translated language. With their deep understanding of language and grammar, the experts

Table 4: The summary of methodology for MT shared-task Lao-Vietnamese

No	Teams	Methodology
1	BGSV AI	Using pre-trained model T5 and Sentence Piece to create a distinct vocabulary set adapted the T5 to train it from scratch for the machine translation task.
2	TESTLAV100	Using OpenNMT and Sentencepiece to tokenizer, apply backtranslation and model weight averaging to optimize performance.
3	FAIZ AIO	Using Transformer wmt en-de big model for Vietnamese to Lao machine translation.
4	BLUESKY	Pre-trained mBART model based on monolingual Vietnamese and Lao for MT task.
5	MTA AI	Using pre-trained models T5, m2m100, and effective backtranslation methods to address the limitations of the Vietnamese-Lao bilingual data. Using M2M-100 418M model and mt5-small for finetuning system.
6	HUMBLE BEES	No technical report.
7	TTS66	Using transformer models, experiment with recurrent neural networks (RNN), optimizing hyper-parameters, and tagged backtranslation.

Table 5: Lao to Vietnamese machine translation task (Automatic evaluation results)

No	Team Name	SacreBLEU	Human	FinalScore
1	BGSV AI	32.56	27.51	27.51
2	TESTLAV100	9.92	12.05	12.05
3	FAIZ AIO	42.19	47.83	47.83
4	BLUESKY	49.67	54.28	54.28
5	HUMBLE BEES	28.45	26.94	26.94
6	TTS66	17.92	33.73	33.73
7	MTA AI	26.03	51.03	51.03

Table 6: Vietnamese to Lao machine translation task (Automatic evaluation results)

No	Team Name	SacreBLEU	Human	FinalScore
1	BGSV AI	29.21	31.56	31.56
2	TESTLAV100	28.09	20.41	20.41
3	FAIZ_AIO	38.04	46.51	46.51
4	BLUESKY	43.08	51.37	51.37
5	MTA_AI	41.88	61.31	61.31

provided feedback and evaluations that closely reflect everyday language use. Employing human evaluation in this manner offers us a proactive and insightful perspective on the model's actual performance, ensuring that the final results are not only technically accurate

but also appropriate and easily understandable for users.

VLSP 2022- MT Chinese-Vietnamese: The human evaluation process was carried out on the translations generated by the models. These tasks included translating Vietnamese into Chinese (Vi-Zh) and Chinese to Vietnamese (Zh-Vi). During this evaluation, translations produced by the models were reviewed and assessed by linguistic experts with extensive skills and knowledge in both languages involved. The objective was to determine the quality, accuracy, and naturalness of the translations to evaluate the performance of the machine translation models.

VLSP 2023-MT Lao-Vietnamese: Human evaluation was carried out on primary runs submitted by participants to two of the MT tasks, namely the MT Vietnamese-Lao (Vi-Lo) task and MT Lao -Vietnamese (Lo-Vi) task.

From the point of view of the evaluation campaign, our goal is to adopt a human evaluation framework able to maximize the benefit for the research community, both in terms of information about MT systems and data and resources to be reused. With respect to other types of human assessment, such as judgments of translation quality (i.e. adequacy/fluency and ranking tasks), the post-editing task has the double advantage of producing (i) a set of edits pointing to specific translation errors, and (ii) a set of additional reference translations. Both these byproducts are very useful for MT system development and evaluation. The human evaluation dataset and the collected post-edits are described in next Section whereas the results of the evaluation are presented in result table.

Evaluation Dataset

The human evaluation datasets each consist of approximately 1,000 sentences, drawn from subsets of the private test sets for each translation task. Specifically, we selected 1,000 sentences for the Zh-Vi and Vi-Zh datasets, and another 1,000 sentences for the Lo-Vi and Vi-Lo datasets. This approach, which involves selecting a consecutive block of sentences for each dataset, was guided by the need to realistically simulate a caption post-editing task.

We received five submissions for each of the Zh-Vi, Vi-Zh, Vi-Lo, and Lo-Vi tasks. For each task, the output from the five systems was given to five professional translators for post-editing on the human evaluation set.

To cope with translators' variability, an equal number of outputs from each MT system was assigned randomly to each translator. The resulting evaluation data for each task consist of the new reference translations for each of the sentences in the human evaluation set. Each one of these references represents the targeted translation of the system output from which it was derived, and remaining additional translations are available as well for the evaluation of each MT system.

Human evaluation results

The outcomes of the two previous rounds of human evaluation through post-editing demonstrated that human evaluation error computed against all the references produced by all post-editors allow a more reliable and consistent evaluation of MT systems with respect to human evaluation error calculated against the targeted reference only. In light of these findings, also this year systems were officially ranked according to human evaluation error calculated on all the collected postedits. In Figure 7 shows the example evaluation by human for Lao - Vietnamese machine translation task (evaluation with the expert in Lao language).

In VLSP 2022, the official results and rankings are presented in bold in Tables 7 and 8, which also present human evaluation error scores calculated on the targeted reference only

	Source	Translation	Score
981	ຕໍ່ມາໄວໜຸ່ມໄດ້ໃຊ້ຄຳວ່າ “ດເຜໄຫມ” ເພື່ອຫມາຍຖຶງການກະທຳມ່ວນຊື່ນ.	Về sau, người trẻ đã dùng từ “lời đề nghị” để nói về những hành động vui vẻ.	50
982	ເຂົ້າໃຈໃນຄວາມຫມາຍກວ້າງກວ່າ, ຜ້າໃຫມ່ຫມາຍຖຶງການສະດວກສະບາຍ ແລະ ເສລີ.	Hiểu được về ý nghĩa rộng hơn, lụa là tự do biểu hiện và tự do.	35
983	Toang ຫມາຍຖຶງມັນຈົບແລ້ວ, ມັນຈົບແລ້ວແທ້ໆ, ໄວໜຸ່ມຍຸກນີ້ໃຊ້ຄຳຫມາຍຖຶງຮ້ອງໄຫ້ເປັນ	Nó đề cập đến nó kết thúc, nó thật sự kết thúc, những năm này ngày nay dùng có nghĩa là những	10
984	ນອກຈາກນັ້ນ, ຈຳນວນຄຳໃຫມ່ແລະປະໂຫຍກທີ່ມີທ່າອ່ຽງໃຫມ່ກໍປະກົດຂຶ້ນ.	Thêm vào đó, có thể có một số từ mới và các câu chuyện mới xuất hiện.	5
985	ຫຼືອາດຈະຖືກຕັ້ງໃຈສະກົດຜິດຊ້ຳນ: ດ້ວກຮຸນເວີ, ເພື່ອຟັງ, ເບິ່ງຄືວ່າຫນ້າຮັກ.	Hoặc có thể bị chọn một luật sai như các cậu bé nói chuyện để nghe, để có vẻ xinh đẹp.	2
986	ອີງຕາມວັດຈະນານຸກົມຫວຽດນາມ (1986) ຂອງສະຖາບັນພາສາຫວຽດນາມ: "ຄຳສະແດງແມ່ນວິທີ	Theo Viện Hàn Quốc (Hình ảnh của Viện Ngôn ngữ Việt Nam) (Hình ảnh của Viện Ngôn ngữ Việt	3
987	ຄຳສະແດງຕະຫລົກຕ່າງຈາກພາສາທາງການທີ່ຖືກນຳໃຊ້. ພາສາ ທາງ ການ ແມ່ນ ພາສາ ຂອງ ຄົນ ທົ່ງ	Các trò lễ khác biệt với ngôn ngữ thường được sử dụng. Ngôn ngữ là ngôn ngữ của mọi người, ai	10
988	ຄຳສະແດງແມ່ນໃຊ້ພຽງແຕ່ໃນພາສາເວົ້າ, ບໍ່ຄ່ອຍເປັນພາສາອ້າງອີງ.	Các tiểu giáo dùng chỉ được dùng trong ngôn ngữ nói, ít phân biệt từ viết.	30
989	ໄວໜຸ່ມນັບມື້ນັບມີສຽງຍາວຂຶ້ນ, ເຂົາເຈົ້າຄິດວ່າການໃຊ້ຄຳສະແດງຈະເຮັດໃຫ້ເວົ້າຍາວປະຈຳວັນມີ	Hàng ngày, giới trẻ càng ngày càng có nhiều tiếng nói hơn, họ nghĩ rằng việc dùng tối sẽ làm cho	2
990	ປະຈຸບັນຄຳສະແດງໄດ້ຂະຫຍາຍອອກໄປ, ແຕ່ລະກຸ່ມໃນສັງຄົມກໍມີຄຳສະແດງຂອງຕົນເອງ.	Hiện nay từ giáo lễ mở rộng, mỗi nhóm trong xã hội cũng có từ giáo lễ riêng của họ.	50
991	ໃນປັດຈຸບັນ, ມັນຍັງມີຄວາມເຊື່ອກັນວ່າການໃຊ້ຄຳສະແດງຈະເຮັດໃຫ້ຄົນເບິ່ງອ່ອນກວ່າໄວ.	Hiện nay, cũng có thể tin là việc dùng từ ngữ sẽ làm cho người ta trở nên mềm lẽ hơn.	55
992	ຄຳສະແດງຂອງ 9x, 10x ໃນປັດຈຸບັນມີຄວາມຫຼາກຫຼາຍທີ່ສຸດ.	Câu chuyện của 9x, 10x hiện tại vô cùng đa dạng.	60
993	ຄຳສະແດງ ບໍ່ແມ່ນທັ້ງດີຫຼືບໍ່ດີທັ້ງຫມົດ.	Okay okay okay tôi có được nó	0
994	ດັ່ງທີ່ຮູ້ກັນດີ, ຖ້າໃຊ້ເພື່ອຈຸດປະສົງທີ່ຖືກຕ້ອງ ແລະ ປານກາງ, ຄຳສັບພາສາສະແຜ່ຈະຊ່ວຍໃຫ້ການ	Như bạn có thể biết, nếu được dùng cho một mục đích đúng và trung bình, thì từ ngữ pháp lý sẽ	50
995	ແນວໃດກໍ່ຕາມ, ຖ້າຖືກທາລຸນ ແລະ ນຳໃຊ້ໃນທາງທີ່ຜິດ, ມັນຈະເຮັດໃຫ້ຄຳສັບສະຫຼັດບັນເທີງຕະຫຼ	Tuy nhiên, nếu bị lạm dụng và dùng một cách sai lầm, thì nó sẽ làm hài hước từ ảo thuật.	45
996	ບັນຄົນປານກາງ, ຄ່າເຮັດຫຼາຍເກີນໄປ, ຖືວ່າມັນບັນຫາປັນເທົ່ງເທົ່ານັ້ນ.	là người trung gian, không bao giờ làm quá nhiều, được coi là một ngôn ngữ giải trí.	20
997	ຫ້າມໃຊ້ຫຼາຍເກີນໄປຈົນລືມພາສາແມ່.	Năm người mẹ quá dùng nhiều đến nỗi họ đã quên đi ngôn ngữ của mẹ mình.	0
998	ຂ້ອຍໄດ້ຫຼິ້ນເພງ ຮ໋າ ວັບ TED ຕ໊ອບ ຫນຶ່ງ ທົດສະວັດແລະຂ້ອຍບໍ່ຄ່ອຍໄດ້ຫຼິ້ນເພງ ໃຫມ່ ຂອງຂ້ອຍ	Tôi đã chơi nhạc cho TED gần một thập kỷ và tôi ít khi nào chơi bài hát nào mới của mình.	55
999	ເມື່ອພົບປະກັບ ທ່ານເລຂາທິການໃຫຍ່ອົງປະຊຸມສະຫະປະຊາຊາດ ກ່ຽວກັບການຄ້າ ແລະ ການພັດ	Tiếp xúc với Tổng Thư ký Hội nghị Liên hợp quốc về Thương mại và Phát triển UNCTAD, Phó Thủ	75
1000	ການ ກຳ ຈັດ ຄຳ ສັບແມ່ນ ຮ໋າ ຄັນເພື່ອໃຫ້ແນ່ໃຈວ່າທຸກໆຄົນໃນທີມເຂົ້າໃຈຢ່າງແນ່ນອນວ່າປະໂຫຍ	Loại bỏ thuật ngữ là rất quan trọng để chắc rằng tất cả mọi người trong đội hiểu chính xác nghĩa	70
1001	ພວກເຮົາຄິດວ່າເຄື່ອງວັດແທກເຫລົ່ານັ້ນມີປະໂຫຍດໃນບາງທາງ, ວ່າພວກເຮົາຕ້ອງການທີ່ຈະເຫັນຂໍ	Chúng tôi nghĩ rằng những số liệu đó thật hữu ích theo cách nào đó, rằng chúng tôi muốn thấy n	75
1002			47,83

Figure 7: An example evaluation by human for Lao - Vietnamese machine translation task

Table 7: Human evaluation results for Chinese-Vietnamese MT. Scores are given in percentage (%)

Task	Rank	Team Name	Result
Zh → Vi	1	SDS	74.73
	2	VBD-MT	71.42
	3	S-NLP	68.74
	4	JNLP	65.29
	5	VC-Datamining	64.40
Vi → Zh	1	S-NLP	73.58
	2	VBD-MT	69.19
	3	VC-Datamining	67.80
	4	SDS	67.68
	5	JNLP	67.08

and results, both on the human evaluation error set and on the full test set, calculated against the official reference translation used for automatic evaluation (see Section 4). Due to various reasons, the S-NLP team could not complete the technical report, so we have removed the S-NLP team from the final standings. As you can see in Table 7, For the Vi → Zh task, the top-ranked system (VBD-MT) is significantly better than all the other systems, while VC-Datamining, SDS and JNLP are not different from each other. On the other hand, For the Zh → Vi task, SDS achieve the highest score, followed by VBDMT, JNLP, and VC-Datamining. Finally, we calculate the average score of both tasks to choose the champion team. The “Final Score” column of Table 8 shows that the winning team is SDS, second is VBD-MT, third is JNLP, and fourth is VC-Datamining. However, there is no system that is significantly better than all other systems; the three top-ranking systems (SDS, VBD-MT, JNLP) are significantly better than the bottom-ranking systems (VC-datamining). To conclude, the post-editing task introduced for manual evaluation brought benefit to the VLSP community, and in general to the MT field. Indeed, producing post-edited versions of the participating systems' outputs allowed us to carry out a quite informative evaluation which minimizes the variability of post-editors, who naturally tend to diverge from the post-editing

Table 8: Final ranking results for Chinese-Vietnamese MT. Scores are given in percentage (%).

Rank	Team Name	Zh → Vi	Vi → Zh	Final Score
1	SDS	74.73	67.68	71.27
2	VBD-MT	71.42	69.19	70.30
3	JNLP	65.29	67.08	66.18
4	VC-Datamining	64.40	67.80	65.60

guidelines and personalize their translations. Furthermore, a number of additional reference translations are made available to the community for further development and evaluation of MT systems.

In VLSP 2023: Official results and rankings are presented in bold in Tables 9, which also present human evaluation error scores calculated on the targeted reference only and results, both on the human evaluation error set and on the full test set, calculated against the official reference translation used for automatic evaluation. Due to various reasons, the S-NLP team could not complete the technical report, so we have removed the S-NLP team from the final standings. As you can see in Table 9, for the Vi → Lo task, the top-ranked system (MTA_AI) is significantly better than all the other systems (Bluesky, Faiz_AIO and BGSV_AI). On the other hand, for the Lo → Vi task, Bluesky achieve the highest score, followed by MTA_AI, Faiz_AIO and BGSV_AI. Finally, we calculate the average score of

Table 9: Final ranking results for Lao↔Vietnamese MT. Scores are given in percentage (%)

Rank	Team Name	Lo-Vi	Vi-Lo	Description
1	Bluesky	**54.28**	51.37	SacreBLEU highest
	MTA_AI	51.03	**61.31**	
2	Faiz_AIO	47.83	46.51	No technical report
	BGSV_AI	27.51	31.56	
3	TTS66	33.73	N/A	
	Humble Bees	26.94	N/A	
	TeslaV100	12.05	20.41	

both tasks to choose the champion team. The "Final Score" column of Table 9 shows that the winning team is Bluesky, second is MTA_AI, third is BGSV_AI).

7. TRANSLATION ERROR ANALYSIS

The evaluation of Vietnamese-Chinese and Vietnamese-Lao machine translation results reveals the systems' impressive performance. Nonetheless, certain limitations persist in both cases, requiring targeted improvements to enhance overall translation quality.

Figure 8 highlights issues such as incorrect name translations, missing key information, and inappropriate word choices, underscoring areas for refinement in the Vietnamese-Chinese translation system. Addressing these challenges offers significant opportunities for enhancing translation accuracy and fluency.

Data	Sample
Input (zh)	原因在于澳大利亚决定取消总额400亿美元的向法国采购核潜艇合同，转而与美国和英国开展联合项目。
Translated (vi)	Nguyên nhân là do Australia quyết định hủy hợp đồng mua tàu ngầm hạt nhân trị giá 4 tỷ USD cho Pháp, chuyển sang triển khai dự án chung với Mỹ và Anh.
Post-processed (vi)	Nguyên nhân là do Australia quyết định hủy hợp đồng mua tàu ngầm hạt nhân trị giá 40 tỷ USD cho Pháp, chuyển sang triển khai dự án chung với Mỹ và Anh.
Input (zh)	目前，美国和欧盟都期待在2021年12月1日前达成解决钢铁和铝贸易争端的协议。
Translated (vi)	Hiện cả Mỹ và EU đều trông đợi một thỏa thuận giải quyết tranh chấp thương mại thép và nhôm trước ngày 1/1/2021.
Post-processed (vi)	Hiện cả Mỹ và EU đều trông đợi một thỏa thuận giải quyết tranh chấp thương mại thép và nhôm trước ngày 1/12/2021.

Figure 8: Common errors of models in Vietnamese-Chinese translation

Similarly, Figure 9 shows that while the Vietnamese-Lao translation system has achieved promising results, it still faces challenges. Proper handling of names and locations remains difficult, requiring further advancements to improve accuracy. Sentences with foreign-language words reduce translation effectiveness, revealing system inflexibility. The models also struggle with dates and mathematical expressions, showing the need for more reliable handling of such cases.

8. CONCLUSIONS

In this paper, we presented the organization and outcomes of the VLSP MT Evaluation Campaign. The VLSP MT evaluation provides a venue where core technologies for spoken language translation can be evaluated on many different languages and compared not only across research teams but also overtime.

- In VLSP 2022, the evaluation was attended by 5 groups: Samsung SDS R&D Center, Vin BigData, Japan Advanced Institute of Science and Technology, Hanoi University of Science and Technology, and VCCorp.
- In VLSP 2023, the evaluation was attended by 7 groups: UET-VNU, MTA, Viettel, Bosch Global Software Technologies Vietnam, HUST, Fulbright University Vietnam, US-VNUHCM. To honor the local organizer, we added among the offered translation directions also Vietnamese-Chinese and Vietnamese-Lao, which finally attracted several participants.

Finally, professional translators conducted a manual evaluation to assess post-editing needs for machine-generated translations. In future work, we plan to extend the task using pre-trained and Vietnamese large language models to support additional languages, including Chinese, Lao, and Khmer.

Data	Sample
Input (lo)	ນການປາກເວົ້າຍາວ 3 ນາທີຂອງນາງ, ດຣ Laura Trice ສະ ເໜີ ຄວາມຄິດກ່ຽວກັບພະລັງຂອງ 2 ຊົ່ວໂມງ & laquo; ຂໍຂອບໃຈ & laquo; - ເຮັດໃຫ້ມິດຕະພາບຂອງທ່ານເລິກເຊິ່ງ ເພີ່ມທະວີຄວາມເປັນມິດ , ແລະໃຫ້ແນ່ໃຈວ່າຄົນອື່ນຮູ້ວ່າມັນມີຄວາມ ໝາຍ ແນວໃດຕໍ່ທ່ານ. ຈົ່ງຍົກເລີກ.
	ຢູ່ໃນຈຸດຄະແນນ ຮາຣ໌ວິກ ແລະ ຈອນສັນ ຍັງຄົງທີ່ໃນລໍາດັບທີໜຶ່ງ ແລະ ສອງ.
Target (vi)	Trong bài nói dài 3 phút , Tiến sĩ Laura Trice trình bày suy nghĩ về sức mạnh của 2 tiếng " cám ơn " – làm sâu sắc thêm tình bạn bè , thắt chặt tình thân , và để chắc chắn rằng người khác biết họ có ý nghĩa như thế nào với bạn . Hãy thử .
	Trong các bảng điểm, Harvick và Johnson duy trì ở vị trí thứ nhất và thứ hai.
Our model's output	Trong bài nói dài 3 phút, Tiến sĩ Laura Trice trình bày một ý tưởng về sức mạnh của 2 tiếng " cám ơn " – làm sâu sắc thêm tình bạn của các bạn, tăng cường tình bạn của các bạn.
	Trên bảng xếp hạng Harvard Vic và Johnson vẫn ổn định ở vị trí thứ nhất và thứ hai.
Meaning (English)	In her 3-minute long speech, Dr. Laura Trice offered her thoughts on the power of 2 hours & laquo; Thank you & laquo; - Deepen your friendships, increase friendships, and make sure others know how much they mean to you. Cancel it.
	In the points standings, Harvick and Johnson remained first and second.
Input (vi)	Andrea Maisi đã mở tỉ số cho Ý ở phút thứ tư với một quả try.
	Thêm vào đó, tất cả các bãi đỗ xe và lối ra vào sẽ được kiểm soát chặt chẽ để ngăn chặn các hành vi đáng nghi.
Target (lo)	ແອນເດຣຍ ມາຊີ ໄດ້ເປີດການທໍາຄະແນນໃນນາທີທີ່ສີ່ໃຫ້ແກ່ອິຕາລີ .
	ນອກຈາກນັ້ນ ບ່ອນຈອດລົດ ແລະ ທາງເຂົ້າ ຈະຖືກກວດກາຢ່າງໃກ້ຊິດ ສໍາລັບກິດຈະກໍາທີ່ໜ້າສົງໃສ.
Our model's output	Andrea Maisi ໄດ້ເປີດອັດຕາເງິນໃຫ້ອິຕາລີໃນນາທີ 4 ດ້ວຍການທົດລອງ .
	ນອກຈາກນັ້ນ, ບ່ອນຈອດລົດແລະທາງເຂົ້າ ທັງຫມົດຈະຖືກຄວບຄຸມຢ່າງເຂັ້ມງວດເພື່ອປ້ອງກັນພຶດຕິກໍາທີ່ຫນ້າສົງໃສ.
Meaning (English)	Andrea Maisi opened the scoring for Italy in the fourth minute with a try.
	In addition, all parking lots and entrances will be strictly controlled to prevent suspicious behavior.

Figure 9: Common errors of models in Vietnamese-Laos translation

ACKNOWLEDGMENT

This work was supported by the Ministry of Science and Technology of Vietnam (Program KC 4.0, project No. KC-4.0.12/19-25). We acknowledge the VLSP organizers for coordinating the Machine Translation challenge and the NLP-UET Lab, VNU Hanoi, for their contributions.

REFERENCES

[1] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *NIPS*, 2014.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online].

Available: http://arxiv.org/abs/1706.03762

[3] K. Cho, B. van Merrienboer, Çaglar Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Conference on Empirical Methods in Natural Language Processing*, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:5590763

[4] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, and M. Norouzi, "Googleś neural machine translation system: Bridging the gap between human and machine translation," *CoRR*, vol. abs/1609.08144, 2016. [Online]. Available: http://arxiv.org/abs/1609.08144

[5] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, "Opennmt: Open-source toolkit for neural machine translation," *CoRR*, vol. abs/1701.02810, 2017. [Online]. Available: http://arxiv.org/abs/1701.02810

[6] P. Koehn and R. Knowles, "Six challenges for neural machine translation," in *Proceedings of the First Workshop on Neural Machine Translation*. Vancouver: Association for Computational Linguistics, 8 2017, pp. 28–39. [Online]. Available: https://aclanthology.org/W17-3204

[7] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *ArXiv*, vol. abs/1508.07909, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:1114678

[8] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040/

[9] M. Post, "A call for clarity in reporting BLEU scores," pp. 186–191, 10 2018. [Online]. Available: https://aclanthology.org/W18-6319

[10] M. Przybocki, K. Peterson, S. Bronsart, and G. Sanders, "The nist 2008 metrics for machine translation challenge overview, methodology, metrics, and results," *Machine Translation*, vol. 23, pp. 71–103, 09 2009.

[11] A. Agarwal and A. Lavie, "Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output," in *WMT@ACL*, 2008.

[12] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual translation with extensible multilingual pretraining and finetuning," *ArXiv*, vol. abs/2008.00401, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:220936592

[13] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "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 (Demonstrations)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 48–53. [Online]. Available: https://aclanthology.org/N19-4009

[14] M. Ott, S. Edunov, D. Grangier, and M. Auli, "Scaling neural machine translation," in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 1–9. [Online]. Available: https://aclanthology.org/W18-6301

[15] T. Kudo and J. Richardson, "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*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71. [Online]. Available: https://aclanthology.org/D18-2012

[16] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, M. Auli, and A. Joulin, "Beyond english-centric multilingual machine translation," *Journal of Machine Learning Research*, vol. 22, no. 107, pp. 1–48, 2021. [Online]. Available: http://jmlr.org/papers/v22/20-1307.html

[17] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Jun. 2021, pp. 483–498. [Online]. Available: https://aclanthology.org/2021.naacl-main.41/

[18] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[20] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual Denoising Pre-training for Neural Machine Translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 11 2020. [Online]. Available: https://doi.org/10.1162/tacl_a_00343

Received on June 29, 2024
Accepted on June 22, 2025