Serving the Underserved: Leveraging BARTBahnar Language Model for Bahnaric-Vietnamese Translation

Long S. T. Nguyen, Tran T. B. Le, Huong P. N. Nguyen, Quynh T. N. Vo, Phong H. N. Nguyen, Tho T. Quan[*]

URA Research Group, Ho Chi Minh City University of Technology, VNU-HCM

[*]**Correspondence:** qttho@hcmut.edu.vn

Abstract

The Bahnar people, one of Vietnam's ethnic minorities, represent an underserved community with limited access to modern technologies. Developing an effective Bahnaric-Vietnamese translation system is essential for fostering linguistic exchange, preserving cultural heritage, and empowering local communities by bridging communication barriers. With advancements in Artificial Intelligence (AI), Neural Machine Translation (NMT) has achieved remarkable success across various language pairs. However, the low-resource nature of Bahnaric, characterized by data scarcity, vocabulary constraints, and the lack of parallel corpora, poses significant challenges to building an accurate and efficient translation system. To address these challenges, we propose a novel hybrid architecture for Bahnaric-Vietnamese translation, with BARTBahnar as its core language model. BARTBahnar is developed by continually training a pre-trained Vietnamese model, BARTPho, on augmented monolingual Bahnaric data, followed by fine-tuning on bilingual datasets. This transfer learning approach reduces training costs while effectively capturing linguistic similarities between the two languages. Additionally, we implement advanced data augmentation techniques to enrich and diversify training data, further enhancing BARTBahnar's robustness and translation accuracy. Beyond leveraging the language model, our hybrid system integrates rule-based and statistical methods to improve translation quality. Experimental results show substantial improvements on bilingual Bahnaric-Vietnamese datasets, validating the effectiveness of our approach for low-resource translation. To support further research, we open-source our code and related materials at `https://github.com/ura-hcmut/BARTBahnar`.

1 Introduction

The Bahnar people, one of Vietnam's 54 ethnic minorities, account for approximately 0.3% of the country's population. As one of the larger minority groups, they possess a rich cultural heritage reflected in unique traditions, festivals, clothing, cuisine, and, most notably, their distinct Bahnaric languages (Bui et al., 2024). This linguistic diversity is a cornerstone of their identity, necessitating dedicated efforts for preservation and promotion. Recognizing this, the Vietnamese government has implemented various policies to safeguard the cultural and linguistic heritage of ethnic minorities, including the Bahnar people. Language preservation plays a pivotal role in maintaining the identity of ethnic groups worldwide. Facilitating linguistic interaction between the Bahnaric and Vietnamese-speaking communities is essential for fostering cultural exchange, mutual understanding, and the preservation of minority identities. Thus, developing an efficient Bahnaric-Vietnamese machine translation system would significantly enhance communication, granting Vietnamese speakers access to the wealth of Bahnaric cultural texts while enabling deeper cross-cultural interactions.

The rapid advancements in Artificial Intelligence, particularly in *Neural Machine Translation* (NMT), have significantly improved translation quality across various language pairs (Qin, 2022). The introduction of the *Transformer* architecture (Vaswani et al., 2017) and subsequent developments in *Large Language Models* (LLMs) have revolutionized *Natural Language Processing* (NLP) applications, including NMT (Wang et al., 2024). Transformer-based models can be broadly categorized into three primary architectures, namely encoder-only, decoder-only, and encoder-decoder. The *encoder-only* type is primarily designed for powerful understanding tasks, making it unsuitable for non-trivial applications such as NMT. Meanwhile, the *decoder-only* architecture excels in text generation but requires large-scale training datasets and lacks explicit encoder support for source language comprehension, making it suboptimal for NMT (Qorib et al., 2024). Additionally, decoder-only models rely on autoregressive generation, which demands substantial computational resources and extensive parallel corpora—both of which are severely lacking for low-resource languages. In contrast, the *encoder-decoder* architecture is inherently suited for NMT, as the encoder effectively captures the semantic and syntactic structure of the source language, while the decoder generates the corresponding translation. However, despite these advantages, building an encoder-decoder-based NMT system for an extremely low-resource language such as Bahnaric remains highly challenging due to severe data scarcity and vocabulary limitations (Ngo et al., 2019). To the best of our knowledge, no prior research has been conducted on Bahnaric-Vietnamese translation.

Given these challenges, it is crucial to examine the linguistic characteristics of Bahnaric and its relationship to Vietnamese. Both languages belong to the *Austroasiatic* family and are considered low-resource in the linguistic landscape (Alves, 2006). Moreover, as both languages coexist within the same country and share a common historical and cultural background, they exhibit notable syntactic similarities and structural overlaps. Additionally, Bahnaric speakers frequently incorporate Vietnamese loanwords, particularly in cases where native Bahnaric vocabulary lacks equivalents (Bui et al., 2024). These linguistic overlaps serve as critical insights for designing an effective translation system.

To leverage these shared linguistic features, we adopt *BARTPho* (Tran et al., 2022), a pre-trained encoder-decoder language model built upon the *Bidirectional and Auto-Regressive Transformers* (BART) (Lewis et al., 2020) architecture, trained on large-scale Vietnamese corpora. This model effectively captures the linguistic characteristics of Vietnamese, making it a strong foundation for adaptation to Bahnaric. To enhance its ability to model the syntactic and lexical properties of Bahnaric, we continually train BARTPho on augmented monolingual Bahnaric data. The model is then fine-tuned on an augmented bilingual Bahnaric-Vietnamese dataset, producing an optimized translation system, which we refer to as *BARTBahnar*. To address the issue of data scarcity in Bahnaric, we implement various *Data Augmentation* (DA) techniques (Li et al., 2022) specifically designed for NMT. These techniques enrich and diversify the training data, improving translation performance. Furthermore, to fully exploit the unique linguistic characteristics of Bahnaric, we propose a novel hybrid approach that integrates BARTBahnar with rule-based and statistical methods. This hybrid strategy enhances translation reliability, particularly in handling loanwords and resolving cases where direct model-generated translations may be inaccurate, ultimately improving translation quality and supporting linguistic preservation.

Our key contributions are summarized as follows.

- We introduced BARTBahnar, an encoder-decoder language model fine-tuned for Bahnaric-Vietnamese translation, leveraging *transfer learning* from BARTPho and various DA techniques. This approach significantly reduces training costs while effectively utilizing linguistic similarities between the two languages to enhance translation performance.

- We designed a robust hybrid system that integrates BARTBahnar with rule-based and statistical methods, effectively handling loanwords and improving translation accuracy.

- We achieved promising translation results on bilingual Bahnaric-Vietnamese datasets, demonstrating the effectiveness of our approach in preserving linguistic heritage and fostering cultural exchange within underserved communities.

2 Related Works

2.1 NMT for Low-resource Languages

NMT has emerged as the dominant paradigm in machine translation, leveraging deep learning models to achieve state-of-the-art performance. However, its reliance on large-scale parallel corpora poses significant challenges for low-resource languages. Existing works addressing these limitations can be broadly categorized into three primary directions: utilizing monolingual data, auxiliary languages, and multi-modal data (Wang et al., 2021).

Monolingual Data Monolingual data, being more abundant and easier to collect than parallel corpora, serves as a critical resource for NMT in low-resource scenarios. Key methodologies include: *(1) Back and Forward Translation*, where pseudo-parallel data is generated by translating monolingual sentences in reverse or the same direction (Sennrich et al., 2016), *(2) Joint Training*, which leverages monolingual data from both source and target languages simultaneously (He et al., 2016), *(3) Unsupervised NMT*, relying on bilingual alignment and iterative back translation (Lample et al., 2018), and *(4) Language Model Pre-training*, where self-supervised training on monolingual data, as demonstrated by models like (Hwang and Jeong, 2023), significantly boosts translation performance. Although these methods effectively exploit monolingual corpora, they heavily depend on high-quality data and often struggle with linguistically distant language pairs, limiting their generalizability.

Auxiliary Languages Closely related languages can facilitate knowledge transfer in low-resource scenarios. Common strategies include: *(1) Multilingual Training*, which shares parameters across multiple language pairs (Johnson et al., 2017), *(2) Transfer Learning*, where models pre-trained on high-resource languages are fine-tuned for low-resource settings (Hujon et al., 2023), and *(3) Pivot Translation*, using an intermediate language to create pseudo-parallel corpora or to combine source-pivot and pivot-target models (Cheng et al., 2017). While these methods leverage linguistic similarities effectively, their success is sensitive to the choice of auxiliary languages, data balancing, and error propagation in pivot-based setups. Moreover, multilingual training can be computationally demanding, posing challenges in resource-constrained contexts.

Multi-modal Data Multi-modal data, such as images and speech, expands the capabilities of NMT by integrating non-textual information. Techniques include: *(1) Image Data*, where image captions generate pseudo-parallel corpora or image features are incorporated into NMT models (Chen et al., 2019), and *(2) Speech-Text Pairs*, supporting translation for languages without written scripts (Zhang et al., 2021). While multi-modal approaches provide valuable support for languages with limited textual resources, they rely on high-quality aligned datasets and face inherent complexity in fusing diverse modalities.

Despite these advancements, most approaches still require large, high-quality datasets, whether monolingual or bilingual, which are unavailable for extremely low-resource languages like Bahnaric. This highlights the critical role of DA techniques in improving NMT performance for low-resource languages.

2.2 Data Augmentation in NMT

To alleviate data scarcity in low-resource NMT, extensive research has focused on DA, which can be grouped into three categories, namely paraphrasing-based methods, noising-based methods, and sampling-based methods (Li et al., 2022).

Paraphrasing-based Methods These methods generate augmented data by altering the original text at lexical, phrase, or sentence levels. For instance, tools like WordNet (Miller, 1994) replace words with synonyms, while *Easy Data Augmentation* (EDA) (Wei and Zou, 2019) offers simple substitution-based strategies. More advanced techniques utilize word embeddings (Wang and Yang, 2015) for enhanced semantic consistency. Although these approaches increase data diversity, they often struggle with preserving sentence meaning, especially in languages with limited lexical resources.

Noising-based Methods These approaches introduce random changes to the original data without maintaining semantic fidelity. Word swapping (Wei and Zou, 2019), sentence-level swapping (Yan et al., 2019), and insertion/deletion (Wei and Zou, 2019) are common examples. While easy to implement, these methods risk disrupting sentence coherence and may be unsuitable for languages with complex syntactic structures.

Sampling-based Methods These methods typically require task-specific knowledge or annotations, such as altering grammatical structures (e.g., converting active to passive voice) (Min et al., 2020) or constructing pseudo-parallel sentences (Zhang et al., 2020). Although effective in generating richer training data, they demand substantial linguistic resources, which are rarely available for extreme low-resource languages.

While DA can significantly boost translation performance, its efficacy for Bahnaric, where grammatical and semantic resources are scarce, remains unknown. Grammar-based approaches can be particularly challenging and may reduce translation accuracy if applied without a deep understanding of the language's structure.

Another standout DA technique is back-translation, which generates entirely new sentences by translating target sentences back into the source language, thus enriching data diversity. For example, (Fabbri et al., 2021) use English-French models to augment French data before training

French-English NMT. Moreover, with the rise of LLMs, (Mai and Luong, 2023) apply GPT-3.5 to augment Vietnamese data, achieving notable improvements in NLP tasks. Nevertheless, deploying back-translation or LLM-based methods for languages like Bahnaric remains challenging due to the lack of pre-trained models and high-quality parallel corpora.

3 Proposed Hybrid Architecture for Bahnaric-Vietnamese NMT

3.1 Overall Pipeline

We propose a comprehensive hybrid system for Bahnaric-Vietnamese translation, consisting of five main phases: Loanword Detection, Word Segmentation, Lexical Mapping, BARTBahnar Translation, and Post-Processing, as illustrated in Figure 1.

The pipeline begins with *Loanword Detection*, which identifies and extracts shared loanwords that appear in both Bahnaric and Vietnamese. These words do not require translation and are excluded from further processing. The remaining words are passed to the *Word Segmentation* phase, where Bahnaric sentences are segmented into meaningful phrases using statistical methods. The segmented phrases are then mapped to their Vietnamese equivalents in the *Lexical Mapping* phase via a bilingual dictionary. Words and phrases that cannot be mapped directly are handled by *BARTBahnar*, which generates Vietnamese translations for the remaining content. Finally, the translated output undergoes *Post-Processing*, ensuring proper sentence structure, punctuation, and grammatical refinements to enhance fluency and accuracy.

3.1.1 Loanword Detection

Loanword detection plays a crucial role in improving translation efficiency by identifying words that are shared between Bahnaric and Vietnamese. This module employs rule-based methods to filter out punctuation marks, special symbols, and numeric characters. Additionally, we utilize a *Named Entity Recognition* (NER) model from a well-established open-source Vietnamese NLP toolkit to detect proper nouns, such as place names and personal names. The identified loanwords are excluded from further translation and directly transferred to the output.

3.1.2 Word Segmentation

As Bahnaric lacks explicit word boundaries, statistical segmentation is necessary to split sentences into meaningful phrases. To construct a phrase dictionary from our monolingual Bahnaric corpus, we employ *Pointwise Mutual Information* (PMI) (Roussinov et al., 2007), a statistical measure that quantifies the strength of association between words. Given an n-gram $(x_1, x_2, .., x_n)$ and $\mathcal{X}^n$ as the set of all possible n-grams extracted from the corpus, the PMI score is computed as shown in Equation 1.

$$\text{PMI}(x_1, x_2, .., x_n) = \log_2 \left(\frac{P(x_1, x_2, .., x_n)}{\prod_{i=1}^{n} P(x_i)} \right), \tag{1}$$

where

$$P(x_1, x_2, .., x_n) = \frac{\text{count}(x_1, x_2, .., x_n)}{\sum_{(x_1, x_2, .., x_n) \in \mathcal{X}^n} \text{count}(x_1, x_2, .., x_n)},$$

$$P(x_i) = \frac{\text{count}(x_i)}{\sum_{x_i \in \mathcal{X}^1} \text{count}(x_i)}.$$

A higher PMI value indicates a stronger association between words, suggesting that they are more likely to form a valid phrase. An n-gram is considered a valid phrase if it satisfies both a minimum frequency threshold and a minimum PMI threshold, as defined in Equation 2 and Equation 3.

$$\text{count}(x_1, x_2, .., x_n) \geqslant \text{min\_freq}, \tag{2}$$

$$\text{PMI}(x_1, x_2, .., x_n) \geqslant \text{min\_pmi}. \tag{3}$$

All valid n-grams are stored in the phrase dictionary. The Bahnaric input is then segmented into phrase units based on this dictionary, facilitating accurate lexical mapping and translation.

3.1.3 Lexical Mapping

This phase employs a bilingual Bahnaric-Vietnamese dictionary to map commonly used words and phrases to their corresponding Vietnamese translations. To efficiently retrieve the most relevant Vietnamese equivalents, we index the dictionary using *Solr* (Tahiliani and Bansal, 2018), an open-source search engine optimized for fast lookup operations. Segments that can be directly mapped are substituted with their Vietnamese counterparts, while unmapped segments are passed to the next translation phase using BARTBahnar.

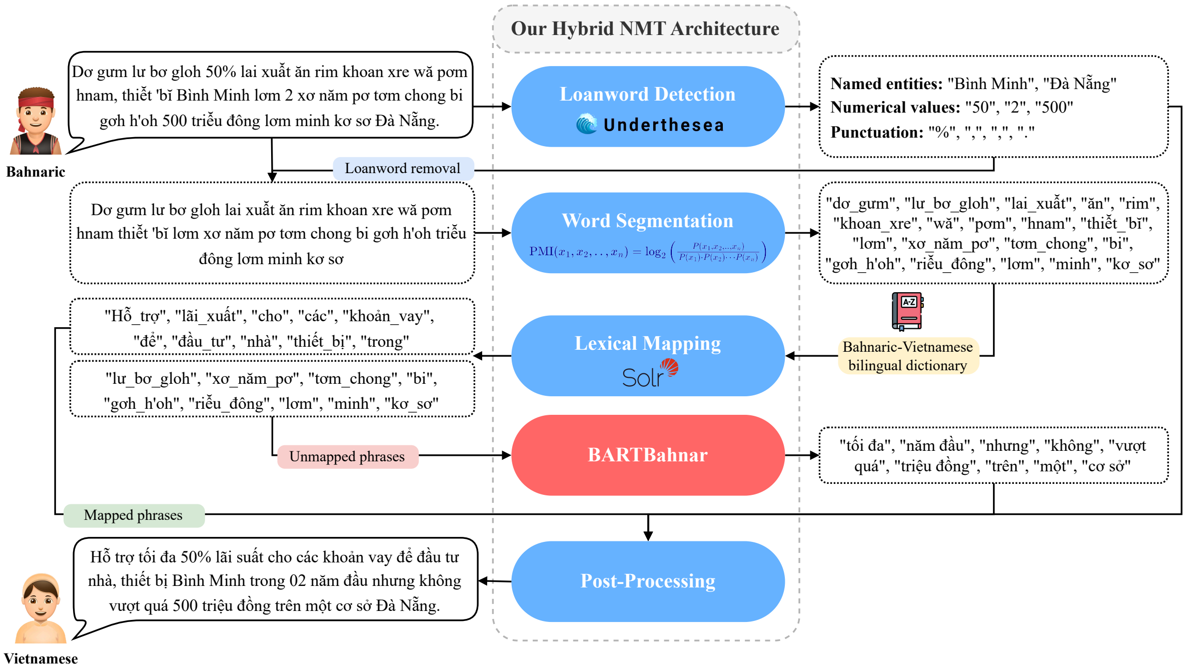
Our Hybrid NMT Architecture
Dơ gưm lư bơ gloh 50% lai xuất ăn rim khoan xre wă pơm hnam, thiết 'bĭ Bình Minh lơm 2 xơ năm pơ tơm chong bi gơh h'oh 500 triễu đông lơm minh kơ sơ Đà Nẵng.
Bahnaric
Loanword Detection
Underthesea
Named entities: "Bình Minh", "Đà Nẵng"
Numerical values: "50", "2", "500"
Punctuation: "%", ",", ",", "."
Loanword removal
Dơ gưm lư bơ gloh lai xuất ăn rim khoan xre wă pơm hnam thiết 'bĭ lơm xơ năm pơ tơm chong bi gơh h'oh triễu đông lơm minh kơ sơ
Word Segmentation
$\mathrm{PMI}(x_1, x_2, \ldots, x_n) = \log_2 \left(\frac{P(x_1, x_2, \ldots, x_n)}{P(x_1) \cdot P(x_2) \cdots P(x_n)} \right)$
"dơ_gưm", "lư_bơ_gloh", "lai_xuất", "ăn", "rim", "khoan_xre", "wă", "pơm", "hnam", "thiết_'bĭ", "lơm", "xơ_năm_pơ", "tơm_chong", "bi", "gơh_h'oh", "riễu_đông", "lơm", "minh", "kơ_sơ"
Bahnaric-Vietnamese bilingual dictionary
Lexical Mapping
Solr
"Hỗ_trợ", "lãi_xuất", "cho", "các", "khoản_vay", "để", "đầu_tư", "nhà", "thiết_bị", "trong"
"lư_bơ_gloh", "xơ_năm_po", "tơm_chong", "bi", "gơh_h'oh", "riễu_đông", "lơm", "minh", "kơ_sơ"
Unmapped phrases
BARTBahnar
"tối đa", "năm đầu", "nhưng", "không", "vượt quá", "triệu đồng", "trên", "một", "cơ sở"
Mapped phrases
Post-Processing
Hỗ trợ tối đa 50% lãi suất cho các khoản vay để đầu tư nhà, thiết bị Bình Minh trong 02 năm đầu nhưng không vượt quá 500 triệu đồng trên một cơ sở Đà Nẵng.
Vietnamese

Figure 1: Illustration of our hybrid NMT architecture, integrating BARTBahnar with rule-based and statistical components. The figure outlines the step-by-step translation process from Bahnaric to Vietnamese. For reference, the English equivalent of the original Bahnaric sentence is *"Support up to 50% interest rate for loans to invest in housing and Binh Minh equipment for the first two years, but not exceeding 500 million VND per facility in Da Nang."*.

3.1.4 BARTBahnar Translation

Unmapped segments that lack direct dictionary translations are processed by *BARTBahnar*, our encoder-decoder language model fine-tuned for Bahnaric-Vietnamese translation. The details of BARTBahnar are elaborated in Section 3.2.

3.1.5 Post-Processing

A critical challenge in *Lexical Mapping* is ambiguity, where multiple Vietnamese candidates may correspond to a single Bahnaric phrase. To resolve this, we implement a scoring mechanism that selects the most contextually appropriate translation, as formulated in Equation 4.

$$v_c = \underset{v_c \in \{v_{c_1}, v_{c_2}, .., v_{c_k}\}}{\operatorname{argmax}} \operatorname{Score}(y_{\text{partial}}, v_c), \quad (4)$$

where v_c is the chosen translation candidate, y_{partial} represents the current state of the translated sentence, and $\operatorname{Score}(y_{\text{partial}}, v_c)$ is computed using a pre-trained language model to ensure fluency and semantic coherence.

After resolving ambiguities, the post-processing module further standardizes punctuation, capitalization, and word order, producing the final Vietnamese translation and completing the pipeline.

3.2 Our BARTBahnar Language Model

We propose a training strategy to effectively adapt a pre-trained language model for low-resource translation, with a specific focus on Bahnaric-Vietnamese. Our approach builds upon *BART*, a sequence-to-sequence model trained as a denoising autoencoder (Lewis et al., 2020), which enhances its ability to reconstruct text under noisy conditions. The model employs a *Bidirectional Encoder* for richer contextual understanding and an *Autoregressive Decoder* for coherent text generation. During training, a random subset of tokens is masked, and the model must autoregressively recover the original sequence, as illustrated in Figure 2.

Our training strategy comprises three main phases: *(1) Pre-training on monolingual Vietnamese data* to capture Vietnamese linguistic features, *(2) Continual pre-training on monolingual Bahnaric data* to adapt the model to Bahnaric text, and *(3) Fine-tuning on bilingual Bahnaric-Vietnamese datasets* for the translation task.

3.2.1 Pre-training on Vietnamese Language

To leverage prior knowledge from a closely related language, we utilize *BARTPho*, a BART model pre-trained on 145 million word-segmented Viet-

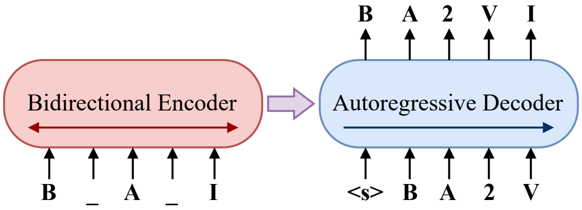

Figure 2: The architecture of BART and its training process as an autoregressive masked language model.

namese sentences. During this stage, two types of noise are introduced, namely *random token masking* and *sentence shuffling*, to enhance the model's ability to handle diverse syntactic structures. Having already learned fundamental properties of Vietnamese grammar and syntax, BARTPho provides a robust foundation for further adaptation to Bahnaric.

3.2.2 Continual Pre-training on Bahnaric Language

We further adapt BARTPho to Bahnaric by training it on a monolingual Bahnaric corpus using an autoregressive *Masked Language Modeling* (MLM) objective, similar to the original pre-training approach. Since Bahnaric is an extremely low-resource language, constructing a high-quality dataset poses a significant challenge.

To address this, we conducted extensive field surveys in Bahnar-speaking regions across Vietnam to gather rare but valuable linguistic materials. Our data sources include: *(1) Direct interviews* with native Bahnar speakers for documenting grammar and vocabulary, *(2) Printed texts* such as religious books, newspapers, song lyrics, and *(3) Local news* bulletins and historical documents. After digitizing and cleaning these materials, we employed a team of annotators to normalize the content, creating a high-quality bilingual Bahnaric-Vietnamese dataset (referred to as the *Original* dataset). Additionally, we applied back-translation techniques to augment this dataset by reconstructing synthetic Bahnaric text from high-quality Vietnamese sentences obtained from *Vietnamese Wikipedia*, leveraging an existing Vietnamese-Bahnaric translation model (Vo et al., 2024). The final dataset statistics are summarized in Table 1.

In this phase, we use only the monolingual Bahnaric portion of the dataset to allow the model to effectively learn Bahnaric syntax and semantics.

Table 1: Statistics of our Bahnaric-Vietnamese bilingual dataset.

Data Source	**Sentence Pairs**
Original	53,942
Back-Translation	270,587
Total	324,529

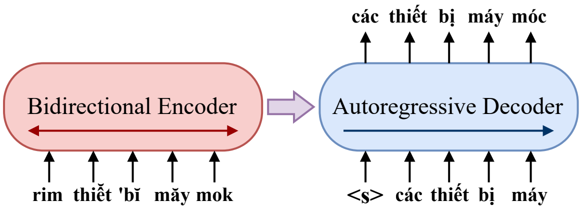

Figure 3: The fine-tuning process for the Bahnaric-Vietnamese translation task.

3.2.3 Fine-tuning for Bahnaric-Vietnamese Translation

After pre-training, we adapt the model for direct translation using a bilingual Bahnaric-Vietnamese dataset. Unlike the MLM phase, where input sequences are partially corrupted, this stage follows a *supervised translation* approach: the encoder takes an unmasked Bahnaric sentence, and the decoder generates the corresponding Vietnamese translation, as shown in Figure 3. During this step, we employ various DA techniques but apply them selectively to the *Original* subset to maintain high-quality supervision, detailed in Section 4.4. We exclude the back-translated data to avoid introducing potential errors, which could otherwise undermine the reliability of the training set.

4 Experimentations

We conduct two main experiments. In the first, we compare our BARTBahnar model against various baselines on the Bahnaric-Vietnamese translation task using only the Original dataset, providing a fair evaluation under limited data conditions. In the second, we examine how different DA techniques affect both BARTBahnar's training process, introduced in Section 3.2.3, and the performance of our end-to-end translation pipeline.

4.1 Dataset

From the Original dataset described in Table 1, we allocate 90% for training and 10% for testing. Although this corpus is relatively small, it is sourced from diverse domains (e.g., economics, social, politics, sports), ensuring a broad range of vocabulary and grammatical constructions.

4.2 Baselines

We select four baselines to compare against BARTBahnar, as described below.

Transformer We replicate the standard Transformer architecture introduced by (Vaswani et al., 2017), following its original hyperparameter configuration.

PhoBERT-Fused NMT Based on (Zhu et al., 2020), we integrate a Bidirectional Encoder into each layer of an encoder-decoder NMT system. In our setup, we replace the baseline's encoder with *PhoBERT*, the encoder component of BARTPho.

ViT5 This is a pretrained Transformer-based encoder-decoder model for Vietnamese (Phan et al., 2022), trained on a large, high-quality Vietnamese corpus using T5-style self-supervision.

BARTPho We employ BARTPho directly, without any Bahnaric-focused continual pre-training.

All baselines are fine-tuned on the Original dataset for 15 epochs with a learning rate of $2e-05$, using the AdamW optimizer (Loshchilov and Hutter, 2019) and hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 1\text{e}-08$.

4.3 Evaluation Metrics

We evaluate translation quality using the *BiLingual Evaluation Understudy* (BLEU) and *Metric for Evaluation of Translation with Explicit ORdering* (METEOR). Both metrics measure lexical and syntactic similarity between the model's output and a reference translation, making them suitable for the Bahnaric-Vietnamese language pair.

4.4 Data Augmentation Methods

Inspired by EDA techniques (Wei and Zou, 2019) and various approaches in the literature, we designed a set of augmentation methods that preserve sentence meaning, maintain grammatical correctness, and introduce controlled variations. This approach balances linguistic diversity with data integrity, ensuring that augmented samples remain useful for training.

Swapping Method Reorders sentence segments within paragraphs or compound sentences, helping the model generalize across varying syntactic patterns.

Combining Method Merges semantically related sentences into more cohesive structures, reducing ambiguities and enriching training examples.

Replacing Method Uses external lexical resources to substitute words with contextually suitable synonyms while preserving semantic consistency. Thematic labels and *part-of-speech* (POS) tagging guide valid replacements.

Insertion and Deletion Methods The insertion method selectively adds thematic words (e.g., locations, time references), providing extra context. The deletion method removes non-essential words, forcing the model to infer missing information and improving robustness against noisy input.

Sliding Window Method Extracts overlapping sub-sequences from sentences, generating samples of varying lengths. By capturing both local and long-range dependencies, it enhances the model's ability to handle diverse input structures.

4.5 Results and Analysis

Table 2 presents the performance of BARTBahnar compared to various baselines on the Bahnaric-Vietnamese translation task. As shown, BARTBahnar consistently outperforms all baselines, validating our transfer learning strategy. By continually pre-training on Bahnaric data, BARTBahnar effectively captures linguistic features from both Vietnamese and Bahnaric, leading to significant improvements in translation accuracy. Notably, the substantial performance drop observed when using BARTPho without Bahnaric-focused continual pre-training demonstrates the necessity of domain adaptation before fine-tuning on the bilingual corpus. These findings reinforce that relying solely on Vietnamese knowledge in BARTPho, even with monolingual Bahnaric training, is insufficient for optimal Bahnaric-Vietnamese translation.

Table 2: Performance comparison of BARTBahnar and baseline models on the Bahnaric-Vietnamese translation task.

Baselines	**BLEU↑**	**METEOR↑**
Transformer	0.26	0.0431
PhoBERT-Fused NMT	2.05	0.2648
ViT5	7.18	0.2386
BARTPho	5.73	0.2076
BARTBahnar	**10.41**	**0.2822**

Beyond baseline comparisons, we also analyze the impact of different data augmentation methods, as shown in Table 3. Notably, the Replacing method, which applies thematic or synonym-based word substitutions, yields the greatest improvements by increasing translation accuracy by up to 200% in certain configurations. This result indicates that broadening vocabulary coverage and introducing controlled lexical variation significantly enhance the model's ability to generalize and capture linguistic nuances in Bahnaric. Additionally, the Deletion method proves effective in this context, since randomly removing words trains the model to handle incomplete source sentences. However, adding excessive noise or distorting sentence structure too much can be counterproductive. For instance, combining Insertion and Swapping leads to a sharp decline in translation quality, likely due to conflicting syntactic cues or disrupted natural sentence formations, thereby undermining model reliability.

Table 3: Effect of various DA methods on our pipeline's translation performance.

DA Methods	**BLEU↑**	**METEOR↑**
Insert + Swap	7.56	0.1905
Insert + Original	12.18	0.2921
Swap	13.74	0.2758
Slide	16.37	0.2640
Combine	16.63	0.3170
Delete	19.45	**0.3323**
Replace (theme)	**20.19**	0.3210
Replace (synonym)	**21.68**	**0.3459**

These results confirm that carefully selecting data augmentation strategies can significantly improve model performance, whereas excessive or poorly suited transformations may introduce noise and reduce accuracy. By strategically applying effective augmentation techniques, particularly synonym replacement, our BARTBahnar-based pipeline achieves better generalization, enhanced robustness, and improved translation quality for Bahnaric-Vietnamese.

5 Conclusion

In this paper, we introduced a novel hybrid architecture for low-resource machine translation, focusing on Bahnaric-Vietnamese and achieving promising results. Alongside rule-based methods that leverage shared features, such as the frequent use of loanwords among Bahnaric speakers to reduce errors and improve translation quality, our key contribution is the custom language model BARTBahnar. This model undergoes a strategic training process: it is first pre-trained on Vietnamese monolingual data, then adapted to Bahnaric monolingual data, and finally fine-tuned for the Bahnaric-Vietnamese translation task. By building on the domestic language model BARTPho, we substantially reduce training costs while relying on structural commonalities between Vietnamese and Bahnaric to maintain high performance. We also investigated various data augmentation methods to identify which techniques are most beneficial for low-resource languages like Bahnaric. Our findings suggest that certain augmentations significantly increase data diversity and enhance translation accuracy, while others may introduce excessive noise, underscoring the importance of carefully selecting augmentation strategies.

Future work could involve further customizing the language model by integrating additional Bahnaric-specific linguistic properties and refining the rule-based components to handle more nuanced text. Exploring additional combinations of data augmentation methods also holds potential for further improvements.

Limitations

Although our system achieves promising results for Bahnaric-Vietnamese translation, several limitations remain. First, it relies on a pre-trained Vietnamese language model, BARTPho, which may not be available for extremely low-resource languages lacking a higher-resource "sibling" language, and training such a model from scratch could be prohibitively expensive. Second, the effectiveness of our transfer learning approach hinges on structural similarities between the two languages; adapting it to languages with drastically different syntax and grammar may pose significant challenges. Finally, the rule-based components in our hybrid system require a bilingual dictionary for phrase mapping, which must be derived from an existing corpus. This can be problematic if the corpus lacks sufficient coverage or quality, and it is labor-intensive to develop in practice.

Acknowledgement

This research is funded by Vietnam Ministry of Science and Technology under the Program "Supporting Research, Development, and Technology Application of Industry 4.0" (KC-4.0/19-25), via the project "Development of a Vietnamese-Bahnaric Machine Translation and Bahnaric Text-to-Speech System (All Dialects)" (KC-4.0-29/19-25).

References

Mark Alves. 2006. Linguistic Research on the Origins of the Vietnamese Language: An Overview. *Journal of Vietnamese Studies*, 1(1-2):104–130.

Long-Ngo-Hoang Bui, Huu-Thien-Phu Nguyen, Minh-Khoi Le, Cong-Thien Pham, and Thanh-Tho Quan. 2024. Handling imbalanced resources and loanwords in Vietnamese-Bahnaric neural machine translation. *International Journal of Intelligent Information and Database Systems*, 16(4):451–472.

Shizhe Chen, Qin Jin, and Jianlong Fu. 2019. From Words to Sentences: A Progressive Learning Approach for Zero-resource Machine Translation with Visual Pivots. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4932–4938. International Joint Conferences on Artificial Intelligence Organization.

Yong Cheng, Qian Yang, Yang Liu, Maosong Sun, and Wei Xu. 2017. Joint Training for Pivot-based Neural Machine Translation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3974–3980.

Alexander Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. 2021. Improving Zero and Few-Shot Abstractive Summarization with Intermediate Fine-tuning and Data Augmentation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 704–717, Online. Association for Computational Linguistics.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 820–828, Red Hook, NY, USA. Curran Associates Inc.

Aiusha V Hujon, Thoudam Doren Singh, and Khwairakpam Amitab. 2023. Transfer Learning Based Neural Machine Translation of English-Khasi on Low-Resource Settings. *Procedia Computer Science*, 218:1–8. International Conference on Machine Learning and Data Engineering.

Soon-Jae Hwang and Chang-Sung Jeong. 2023. Integrating Pre-Trained Language Model into Neural Machine Translation . In *2023 2nd International Conference on Frontiers of Communications, Information System and Data Science (CISDS)*, pages 59–66, Los Alamitos, CA, USA. IEEE Computer Society.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised Machine Translation Using Monolingual Corpora Only. In *International Conference on Learning Representations*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3:71–90.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Hieu-Hien Mai and Ngoc Hoang Luong. 2023. Data Augmentation with GPT-3.5 for Vietnamese Natural Language Inference. In *2023 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 435–440.

George A. Miller. 1994. WordNet: A Lexical Database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic Data Augmentation Increases Robustness to Inference Heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352, Online. Association for Computational Linguistics.

Thi-Vinh Ngo, Thanh-Le Ha, Phuong-Thai Nguyen, and Le-Minh Nguyen. 2019. Overcoming the Rare Word Problem for low-resource language pairs in Neural Machine Translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 207–214, Hong Kong, China. Association for Computational Linguistics.

Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H. Trinh. 2022. ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student*

Research Workshop, pages 136–142, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Mo Qin. 2022. Machine Translation Technology Based on Natural Language Processing. In *2022 European Conference on Natural Language Processing and Information Retrieval (ECNLPIR)*, pages 10–13.

Muhammad Qorib, Geonsik Moon, and Hwee Tou Ng. 2024. Are Decoder-Only Language Models Better than Encoder-Only Language Models in Understanding Word Meaning? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16339–16347, Bangkok, Thailand. Association for Computational Linguistics.

Dmitri Roussinov, SzeWang Fong, and David B. Skillicorn. 2007. Detecting word substitutions: PMI vs. HMM. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 885–886. ACM.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Sahitya Tahiliani and Ayush Bansal. 2018. Comparative Analysis on Big Data Tools: Apache Solr Search and Hibernate Search. In *2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pages 164–170.

Nguyen Luong Tran, Duong Minh Le, and Dat Quoc Nguyen. 2022. BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Hoang Nhat Khang Vo, Duc Dong Le, Tran Minh Dat Phan, Tan Sang Nguyen, Quoc Nguyen Pham, Ngoc Oanh Tran, Quang Duc Nguyen, Tran Minh Hieu Vo, and Tho Quan. 2024. Revitalizing Bahnaric Language through Neural Machine Translation: Challenges, Strategies, and Promising Outcomes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):23360–23368.

Rui Wang, Xu Tan, Renqian Luo, Tao Qin, and Tie-Yan Liu. 2021. A Survey on Low-Resource Neural Machine Translation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4636–4643. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

William Yang Wang and Diyi Yang. 2015. That‘s So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors using #petpeeve Tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563, Lisbon, Portugal. Association for Computational Linguistics.

Yanshu Wang, Jinyi Zhang, Tianrong Shi, Dashuai Deng, Ye Tian, and Tadahiro Matsumoto. 2024. Recent Advances in Interactive Machine Translation With Large Language Models. *IEEE Access*, 12:179353–179382.

Jason Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Ge Yan, Yu Li, Shu Zhang, and Zhenyu Chen. 2019. Data Augmentation for Deep Learning of Judgment Documents. In *Intelligence Science and Big Data Engineering. Big Data and Machine Learning*, pages 232–242, Cham. Springer International Publishing.

Chen Zhang, Xu Tan, Yi Ren, Tao Qin, Kejun Zhang, and Tie-Yan Liu. 2021. UWSpeech: Speech to Speech Translation for Unwritten Languages. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14319–14327.

Yi Zhang, Tao Ge, and Xu Sun. 2020. Parallel Data Augmentation for Formality Style Transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3221–3228, Online. Association for Computational Linguistics.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. 2020. Incorporating BERT into Neural Machine Translation. In *International Conference on Learning Representations*.