Revitalizing Bahnaric Language through Neural Machine Translation: Challenges, Strategies, and Promising Outcomes

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, Tho Quan*

Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, Ho Chi Minh City, Vietnam
{khang.vo872003, dong.le0110, dat.phantran8136, sang.nguyen.imp21, pqnguyen.sdh20, oanh.tranotsc1123, nqduc, hieu.votranminh21, qttho}@hcmut.edu.vn

Abstract

The Bahnar, a minority ethnic group in Vietnam with ancient roots, hold a language of deep cultural and historical significance. The government is prioritizing the preservation and dissemination of Bahnar language through online availability and cross-generational communication. Recent AI advances, including Neural Machine Translation (NMT), have transformed translation with improved accuracy and fluency, fostering language revitalization through learning, communication, and documentation. In particular, NMT enhances accessibility for Bahnar language speakers, making information and content more available. However, translating Vietnamese to Bahnar language faces practical hurdles due to resource limitations, particularly in the case of Bahnar language as an extremely low-resource language. These challenges encompass data scarcity, vocabulary constraints, and a lack of fine-tuning data. To address these, we propose transfer learning from selected pre-trained models to optimize translation quality and computational efficiency, capitalizing on linguistic similarities between Vietnamese and Bahnar language. Concurrently, we apply tailored augmentation strategies to adapt machine translation for the Vietnamese-Bahnar language context. Our approach is validated through superior results on bilingual Vietnamese-Bahnar language datasets when compared to baseline models. By tackling translation challenges, we help revitalize Bahnar language, ensuring information flows freely and the language thrives.

Introduction

The Bahnar people, also known as Ba-Na (pronounced as [ɓaː˧naː˧] in Vietnamese), constitute a unique ethnic minority group within the diverse mosaic of ethnic communities in Vietnam. In contemporary times, there are concerted efforts led by the Vietnamese government aimed at promoting their integration into mainstream society, focusing on aspects like socio-cultural and scientific literacy. As part of this initiative, there is a notable emphasis on translating important documents into the Bahnar language, involving both government authorities and local communities.

Simultaneously, domestic research organizations have shown an increasing interest in developing automated translation systems capable of translating from Vietnamese to various Bahnar language ethnolects. In recent years, the field of Artificial Intelligence (AI) has witnessed remarkable advancements, particularly in the realm of NMT (Tan et al. 2020). These breakthroughs have ushered in a new era of translation, characterized by unprecedented levels of accuracy and fluency. NMT, in particular, has emerged as a powerful tool that not only bridges language barriers but also significantly enhances accessibility for Bahnar language speakers. It has opened up new avenues for them to access information and engage with content that was previously beyond their reach. This technological leap is not just a translation tool; it represents a catalyst for empowerment and inclusivity, bolstering the preservation and revitalization of the Bahnar language and its diverse culture.

However, translating Vietnamese to Bahnar language is a task fraught with practical challenges, primarily stemming from the stark resource limitations that characterize Bahnar language as an extremely low-resource language. These challenges manifest in various forms, including data scarcity and vocabulary constraints. Furthermore, NMT systems based on large and complex deep learning architectures pose challenges when deployed as popular tools to reach the Ba Na people in remote and distant regions of Vietnam.

To tackle those problems, we propose to adopt an approach of transfer learning principles, leveraging carefully selected pre-trained models. By doing so, we aim to not only enhance the quality of translation but also optimize computational efficiency. Moreover, we also harness the inherent linguistic similarities between Vietnamese and Bahnar language, establishing a foundation upon which our translation model can be constructed. To address the resource scarcity issue for the Bahnar language, we integrate specialized augmentation strategies into the translation task. They are designed to adapt and fine-tune our model, ensuring that it performs optimally in translating between these two languages.

Contributions of our work are as follows:

- We present a Vietnamese-Bahnar language translation mechanism that combines the strengths of NMT and the analysis of linguistic similarities between Vietnamese and Bahnar language. Our NMT model is built upon the pre-trained BARTpho architecture, a sequence-to-sequence Transformer-based model fine-tuned on Vietnamese, enabling it to understand Vietnamese language intricacies. We further fine-tune this model using a pro-

[*] Corresponding authors

posed specific strategy on a collectible Bahnar language dataset. The resulting model, known as BN-BARTpho, retains the characteristics of Vietnamese while being capable of generating corresponding Bahnar language translations. We also introduce a pipeline in which BN-BARTpho is incorporated with other processes that exploit the linguistic similarities between Vietnamese and Bahnar language.

- To overcome the resource limitations of Bahnar language, we also introduce appropriate data augmentation methods for that language. In the field of Natural Language Processing (NLP), data augmentation techniques have been proposed primarily for text classification tasks. We also analyze and adapt augmentation techniques suitable for NMT. The combination of BN-BARTpho and augmentation techniques yields a translation model that is able to demonstrate a superior performance compared to other well-known baseline models.

Related Works

Machine Translation System Based on Relatively-Sized Models

Machine translation systems based on deep learning models have been extensively researched for quite some time and have achieved significant milestones with the emergence of techniques such as Word2Vec (Mikolov et al. 2013) for word representation, sequential processing models like Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997), sequence-to-sequence architectures (Sutskever, Vinyals, and Le 2014), and especially the introduction of Attention mechanisms (Dosovitskiy et al. 2021). These advancements have opened up promising avenues for machine translation, natural language processing, and even image processing.

Following a common design comprising two main blocks, one for encoding input information (encoder) and the other for generating translated sentences (decoder), the initial models utilizing this design included two RNNs, or more advanced versions with two LSTMs, allowing for a smooth process of reading, understanding, and translating (Sutskever, Vinyals, and Le 2014). However, RNN models had limitations in ”remembering” large-sized input sequences. The introduction of attention mechanisms like Bahdanau Attention (Bahdanau, Cho, and Bengio 2015) and Luong Attention (Luong, Pham, and Manning 2015) set the stage for subsequent research not only in machine translation but also in other areas like question-answering systems, information retrieval, information extraction, and more.

In the narrower domain of machine translation, recent research has focused on several key directions:

- Improving training speed and prediction efficiency with real-world data.
- Reducing model size for deployment on smaller devices by utilizing shared weight matrices in model design (Takase and Kiyono 2023).
- Training with limited data using transfer learning methods (Zhu et al. 2020) and data augmentation (Zhu et al. 2020; Takase and Kiyono 2021).
- Enhancing model accuracy and fluency. (Xia et al. 2017; Currey, Miceli Barone, and Heafield 2017)

For the Vietnamese language, various translation models between Vietnamese and other languages have been designed, deployed in services like Google Translate, or made available as pre-trained models integrated into libraries like HuggingFace. These models cover seven language pairs: Vietnamese-English, Vietnamese-Esperanto, Vietnamese-French, Vietnamese-Russian, Vietnamese-German, Vietnamese-Spanish, and Vietnamese-Italian. With substantial data, these models achieve relatively high accuracy, with the highest being the Vietnamese-English translation model (42.8 BLEU Score) and the lowest for the Vietnamese-Esperanto pair (12.2 BLEU Score).

Many studies have explored the use of multilingual models, aiming to share a common encoder and separate decoders for each language (Vázquez et al. 2019) to address data scarcity issues. However, the authors of PhoBERT - a Vietnamese pre-trained language model, instead of BERT for better Vietnamese presentation extraction (Nguyen and Nguyen 2020), have demonstrated that using models dedicated to fixed language pairs yields superior results compared to using a single multilingual model. They have introduced the BARTpho model (Tran, Le, and Nguyen 2022) along with pre-trained weights trained on Vietnamese data.

Data Augmentation in NMT

Data augmentation (DA) was first widely applied in the computer vision field and then used in natural language processing, achieving improvements in many tasks. DA help to improve the diversity of training data, thereby helping the model anticipate the unseen factors in testing data. Many DA strategies for NLP, from rule-based manipulations (Zhang, Zhao, and LeCun 2015) to more complex generative systems (Liu et al. 2020), have been developed despite challenges associated with text.

Some auxiliary tasks have been previously used for DA, but mostly on the source side and rarely within a multi-task learning (MTL) framework. For instance, Zhang et al. 2020 applied the technique of replacing tokens with placeholders, specifically in the source language. Similarly, Xie et al. 2017 evaluated the impact of replacements on the target data. Another related approach is word dropout, which has been explored by (Gal and Ghahramani 2016) and (Arora, Liang, and Ma 2017).

In terms of altering word order, there have been several proposals. Artetxe et al. 2017 and Lample et al. 2018 have put forward their respective strategies. However, it is notable to mention the approach suggested by Zhang et al. 2019, which involves a self-translation technique utilizing a right-to-left decoder. Their method requires generating translations from the model during training and making adjustments to multiple terms in the training loss.

There are additional noteworthy DA approaches that involve word replacement. Xie et al. 2017 employ random word replacement on the source side of training samples. Gao et al. 2019 replace randomly selected words with soft words, whose representations are derived from the probability distribution provided by a language model. Wei and Zou has also applied this approach in EDA where they randomly replaced n words with their synonym. (Fadaee, Bisazza, and Monz 2017) replace several words in their training samples with infrequent words to enhance the NMT model's performance when translating such words.

In the context of back-translation, Edunov et al. 2018 experimented with various straightforward transformations such as word deletion, replacement, and swapping on the back-translated data, resulting in a noticeable improvement. In terms of the special token used to prevent negative transfer between tasks, (Caswell, Chelba, and Grangier 2019) propose a similar approach to identify synthetic samples when combining actual parallel data and back-translated data for training.

From empirical research, back-translation and word replacements are the two most common DA techniques for NMT. For back-translation, this technique uses monolingual data to augment a parallel training corpus. Although useful, back-translation is frequently susceptible to mistakes in initial models, a typical issue with self-training algorithms (Chapelle, Scholkopf, and Zien 2009). The second category is based on word replacements. This approach is a feasible option for the Bahnar language because it produces reasonably high-quality augmented data and works best with low-resource datasets. Moreover, the noising method applied to monolingual corpus has shown its effect with the encoder-decoder NMT model. The noising-based methods may not be the dominant approach in NMT; however, they have been researched and applied in low-level architecture during the training process. Although they have proved their potential, very few of them have ever been applied to produce visible results from augmented data like text classification. Inspired by EDA and the multi-task learning approach (Sánchez-Cartagena et al. 2021), a multi-task learning framework with multiple operations should be conducted and examined to assess its effectiveness with individual methods. Besides, a further approach of the nosing-based method on the sentence level should be investigated which is a promising way of applying the noising-based method on low-resource NMT.

Background

Introduction to Machine Translation

The machine translation problem has been tackled using various methods, including rule-based approaches and phrase-based translation. However, in modern approaches, the use of deep learning models and sequence-to-sequence (Seq2Seq) architectures is leading the way in research.

To apply this architecture to the machine translation task, textual data needs to be transformed into a format that the machine can understand, using tokenization techniques where each word is represented as a token and encoded as a dense vector (also known as Word Embedding). These vectors are generated by models like Word2Vec to capture the relationships between words in vector space, as we will discuss later in this paper.

In this architecture, the encoder's role is to extract the meaning of the input sentence, including the meanings of its constituent words and the relationships between words. The decoder, on the other hand, acts as a language generation model, with the vector W encoded by the encoder. With the information c extracted from the input sequence x_i through the encoder, the decoder predicts the next word y_t given c and the previously predicted words $y_0, ..., y_{t'-1}$:

$$p(Y) = \prod_{t=1}^{T} p(y_t|y_1, ..., y_{t-1}, c)$$

Although it was the best model for machine translation in 2014, its main weakness lies in the length of the input and output data. RNN models have the weakness of relying only on a small number of nearby words, while LSTM models, despite using gates to preserve information over multiple steps, still perform well on short sentences and struggle to handle text data.

BLEU Score

BLEU (Bilingual Evaluation Understudy) (Papineni et al. 2002) score is a metric used in machine translation to assess translation quality. It measures the precision of n-grams (consecutive word sequences) in the candidate translation compared to reference translations. A perfect translation receives a BLEU score of 100%. Higher BLEU scores indicate better translation quality, but interpretation depends on context and language.

Transformer Attention

With the same concept of computing dependency scores, the authors (Vaswani et al. 2017) proposed a method for calculating attention scores. By transforming the representation vectors of the input words $X = x_i$ and the target vectors for which attention needs to be computed $Y = y_j$ (self-attention if X and Y belong to the same encoder or decoder side, or regular attention if Y belongs to the decoder side), they project them into a different vector space through transformations $Q = f_q(Y)$, $K = f_k(X)$, and $V = f_v(X)$.

The attention score will be calculated as follows:

$$A = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

The attention vector is the weighted average of the vector set V, weighted by A:

$$O = AV = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Transformer Model and Its Variants

Transformer model. Recognizing the limitations of RNN models, which process sequences sequentially and do not fully exploit parallel computing resources, Vaswani et al. introduced the Transformer model. In this model, the encoder consists of stacked self-attention layers, while the decoder comprises both self-attention and conventional attention layers, augmented with positional encoding. By extracting contextual vectors from the encoder, the Transformer model achieved a BLEU score of 41.0 on the WMT'14 dataset, becoming the state-of-the-art model for machine translation, surpassing contemporary models.

The encoder and decoder in the Transformer model consist of multiple layers. Each encoder layer includes a self-attention layer and a neighboring feed-forward layer. In contrast, each decoder layer adds an attention layer in between, in order to serve the purpose of context extraction from output of the encoder.

Several machine translation models built upon the Transformer framework. MarianMT, a Transformer model with 6 layers in both the encoder and decoder, is integrated into the Huggingface library, offering over 1000 pre-trained weight sets for various language pairs. For Vietnamese, there are a total of 7 weight sets for language pairs such as Vietnamese-English, Vietnamese-Esperanto, Vietnamese-French, Vietnamese-Russian, Vietnamese-German, Vietnamese-Spanish, and Vietnamese-Italian. The recent T5 model (Raffel et al. 2020), developed by Google, uses a Transformer architecture with 12 encoder and decoder layers, making minor structural modifications (Devlin et al. 2019). The BERT-fused model (Zhu et al. 2020) combines BERT into the attention layers of the Transformer model, leveraging monolingual data to enhance translation performance, especially with small bilingual datasets. This fusion provides contextual information for both encoder and decoder sides, improving translation accuracy.

Methodology

Overview of Pipeline

Figure 1 presents the whole process of our Vietnamese-Bahnar language translation system. Bahnar language are considered low-resource languages, meaning that they have limited linguistic data and resources available for natural language processing tasks. Therefore, conventional methods of machine translation that rely on large parallel corpora for training are not feasible or effective for this task. However, Vietnamese and Bahnar language share some common features in their grammatical structure, such as word order, morphology, and syntax. These features can be exploited to facilitate the machine translation process and improve the quality of the output. The proposed method is based on a chunking translation approach that utilizes the combination of word mapping and fine-tuning a pretrained language model for machine translation. The chunking translation approach is less computationally complex than conventional methods that require full sentence alignment and parsing. The proposed method consists of an end-to-end pipeline that has two main phases:

- **Segmentation phase:** This phase takes a Vietnamese sentence as input and applies techniques such as word segmentation and named entity recognition (NER) to identify whether each word or phrase is an anchor or a chunk. Anchors are words or phrases that can be directly mapped to Bahnar language, while chunks are words or phrases that require further translation. And the output of this phase will be a list of anchors and chunks that constitute the input sentence.
- **Mapping phase:** This phase takes the list of anchors and chunks as input and employs different techniques to map them to Bahnar language. For anchors, the mapping method depends on whether they are words existing in the dictionary or entities. For words in the dictionary, a word mapping method is used, which looks up the corresponding Bahnar language word in a bilingual dictionary. For entities, an entity mapping method is employed, which applies phonetic rules to convert the Vietnamese entity into a Bahnar language entity. For chunks, a fine-tuned BARTpho model is used to translate them into Bahnar language phrases. And the output of this phase will be a list of translated segments from anchors and chunks.

After the two phases, the list of translated segments is concatenated together to form a complete Bahnar language sentence. Figure 2 depicts an example of the translation from a Vietnamese sentence to a Bahnar language sentence based on our proposed model pipeline.

Segmentation Phase

In this phase, we use VnCoreNLP (Vu et al. 2018), a toolkit for Vietnamese natural language processing, to perform word segmentation and named entity recognition (NER) on the input sentence.

- *Word segmentation* (Nguyen et al. 2018), which is one of the modules of VnCoreNLP, is employed to split the Vietnamese sentences into words. This is an essential step for Vietnamese, as words are not separated by spaces. Then, each token is checked against a Vietnamese-Bahnar language bilingual dictionary to identify whether it is an anchor or not. If it is an anchor, it can be directly mapped to Bahnar language.
- *Named entity recognition*, a module of VnCoreNLP, is used to identify and label entities in sentences, such as person names, locations, organizations, dates, etc. This preserves the entities as anchors in the translation process and avoids errors caused by word segmentation.
- *Punctuation marks*, numbers are also considered as important anchors for chunking. Punctuation marks, indicating the boundaries of sentences and clauses, can be used to split a sentence into smaller chunks.
- *Chunks* are words or phrases that are not anchors. They are extracted by finding the boundaries between anchors.

Bahnaric-Fine-Tuned BN-BARTpho

BARTpho (Tran, Le, and Nguyen 2022) is a pretrained sequence-to-sequence model for Vietnamese based on the BART architecture (Lewis et al. 2020) that is trained on a large corpus of Vietnamese text. BART is a sequence-to-sequence model that can generate text from text, such as summarization, translation, or text generation. Based on this pre-trained model, we introduced the Bahnaric-fine-tuned BN-BARTpho which further fine-tuned BARTpho with Bahnar language dataset for the downstream of Vietnamese-Bahnar language datasets as follow:

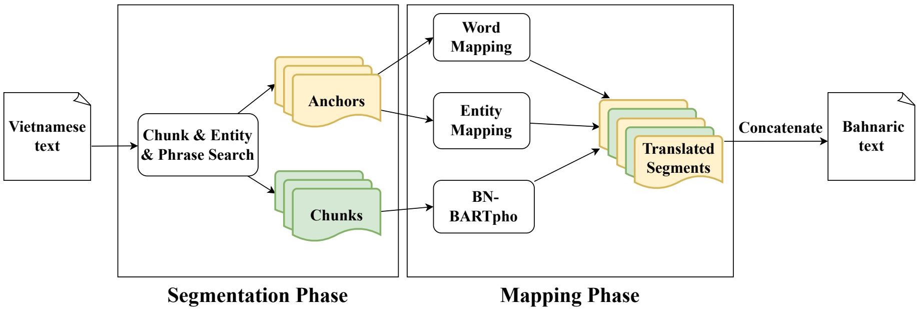
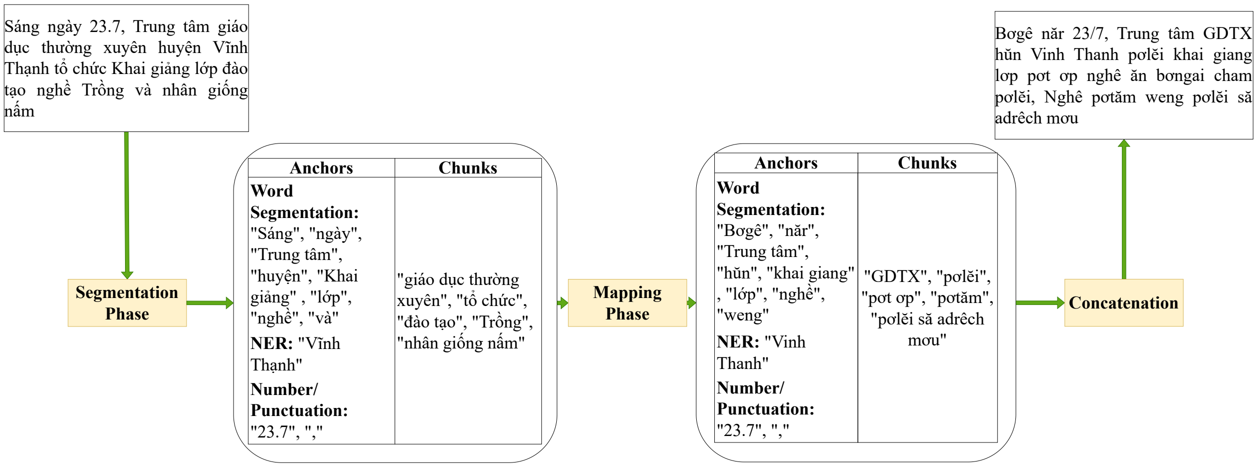


Figure 1: The overall process of Vietnamese-Bahnar language translation



Sáng ngày 23.7, Trung tâm giáo dục thường xuyên huyện Vĩnh Thạnh tổ chức Khai giảng lớp đào tạo nghề Trồng và nhân giống nấm

Anchors	Chunks
Word Segmentation: "Sáng", "ngày", "Trung tâm", "huyện", "Khai giảng" , "lớp", "nghề", "và" **NER:** "Vĩnh Thạnh" **Number/ Punctuation:** "23.7", ","	"giáo dục thường xuyên", "tổ chức", "đào tạo", "Trồng", "nhân giống nấm"

Anchors	Chunks
Word Segmentation: "Bơgê", "năr", "Trung tâm", "hŭn", "khai giang" , "lớp", "nghề", "weng" **NER:** "Vinh Thanh" **Number/ Punctuation:** "23.7", ","	"GDTX", "pơlĕi", "pơt ơp", "pơtăm", "pơlĕi să adrêch mơu"

Bơgê năr 23/7, Trung tâm GDTX hŭn Vinh Thanh pơlĕi khai giang lơp pơt ơp nghê ăn bơngai cham pơlĕi, Nghê pơtăm weng pơlĕi să adrêch mơu

Figure 2: An example of our Vietnamese-Bahnar language translation pipeline

- *Dataset*: The dataset for fine-tuning consists of parallel corpora, bilingual dictionary, anchors, and chunks. Parallel corpora are pairs of sentences in Vietnamese and Bahnar language that are aligned at the sentence level. Bilingual dictionary is a list of words and their translations in both languages. Anchors and chunks are extracted from applying the segmentation phase to the dataset.
- *Training procedure*: To fine-tune BARTpho, we freeze the encoder of the model and only train the decoder. The encoder takes a Vietnamese chunk as input and encodes it into a latent representation. The decoder takes the latent representation and the previous tokens to generate a Bahnar language chunk as output. The model is trained by minimizing the cross-entropy loss between the output and the target chunk to optimize the decoder parameters

Mapping Phase

We perform the following steps to map the anchors, chunks to Bahnar language segments.

Anchors: We examine each anchor and compare it with the bilingual dictionary. Depending on the type of anchor, we apply different mapping rules:

- *Word*: It is mapped to the corresponding word in Bahnar language that is listed in the dictionary.
- *Entity*: Here, it is normalized just by removing the tone mark of Vietnamese.
- *Punctuation mark or number*: It is kept unchanged.

Chunks: The fine-tuned BN-BARTpho is employed to translate the chunks from Vietnamese to Bahnar language. We feed each chunk as an input to BN-BARTpho and obtain an output chunk in Bahnar language.

Data Augmentation Methods

Multi-task learning data augmentation (MTL DA) This approach is motivated by two studies: EDA and DA operations from multi-task learning data augmentation framework (Sánchez-Cartagena et al. 2021). The method combines a set of simple DA transformations to produce synthetic target sentences to strengthen the encoder. Some brief explanations of each auxiliary task are presented below. Certain transformations are regulated by a hyperparameter α, which determines the percentage of target words influenced by the transformation. With t is the number of tokens in our target sentences, we have the following methods:

- **Swap:** Randomly swapping $\alpha \cdot t$ target words
- **Token:** Randomly replacing $\alpha \cdot t$ target words with a special (UNK) token
- **Source:** Copying the source sentence
- **Reverse:** Reversing the word order of the target sentence

- **Replace:** Randomly substituting $\alpha \cdot t$ target words and their corresponding aligned source words

Sentence Boundary Augmentation This approach was originally motivated by the proposal of Li et al. 2021. The method divided the target sentence into two parts. For two adjacent sentences, their parts will be swapped and combined with the other parts. Moreover, Luque 2019 has stated a significant point that the swapping method on the sentence level could still keep the semantic meaning. Therefore, sentence boundary augmentation has been applied in the context of low-resource translation, and it has also exposed the model to bad segmentation during training.

Experiments

Dataset

Our research relied heavily on extensive fieldwork, as collecting reliable data is not an easy task at all. While these communities thrive across three primary regions – Binh Dinh, Gia Lai, and Kon Tum – researchers decided to contact Department of Science and Technology - People's Committee of Binh Dinh Province, in order to request collecting data from Binh Dinh local people, as its linguistic unity extends beyond geography; Bahnar language dialects spoken in Binh Dinh remain readily comprehensible to those in Gia Lai and Kon Tum (this is guaranteed both by local people and by researchers ourselves analyzing semantic as well as sentence structure). Choosing Binh Dinh as the primary data field thus felt like a natural choice – not only would it grant us deep access to the core of Bahnar language culture, but its central location and shared dialects allowed us to effectively capture the essence of the broader Bahnar language experience. Our dataset includes:

- Bahnar language handwritten words of Binh Dinh area (collected from Vietnamese - Bahnar language bilingual dictionary. Through a dedicated process of transcription, the handwritten data is converted into digital format, thereby laying the foundation for effective data management and future research endeavors). We mainly gather data in Binh Dinh area because it is the place where the majority of Bahnar language people are living.
- From voice recordings of Vinh Thanh (a district of Binh Dinh) radio station, we have some local people who are fluent in both Vietnamese and Bahnar language translating those voice records and documenting them into sentences.
- Documents about scientific and technological, socio-economics, political, social, cultural, sports,... written in Binh Dinh Bahnar language.
- A Binh Dinh Bahnar language dialect handbook, collected from local people.

Our collected data are common sentences (by having local people written down some daily basic sentences that they use everyday. The idea is to collect the most common and useful greetings used by the Bahnar people, the same way we learn the basics ”Hello, how are you?” and ”I'm fine, thank you” if we learn English), formal and informal conversations, narrative stories, and folktales written in Binh Dinh Bahnar language. The dataset was divided into three sub-datasets: A training set, a test set, and a validation set, which were used for training, testing, and validating, respectively. In total, our training set contains a set of $16,105$ pairs of sentences, the test set contains $1,988$ pairs, and the valid set has a set of $1,987$ pairs of sentences.

Experiment Setup

Augmenting dataset The training dataset were expanded by augmentation, resulting in a final training dataset that is twice the size of the original. As the augmented data is evaluated across multiple models, the augmentation process is performed iteratively, generating a new training dataset for each evaluation. This repetitive augmentation is designed to assess the proposed models' ability to adapt to variations within the dataset, providing insights into their robustness and adaptability to diverse linguistic instances in the low-resource Bahnar language.

In the augmentation process, our choice of hyperparameter α value at 0.5 serves as a deliberate choice to strike a balanced augmentation approach. With an alpha value of 0.5, the augmentation process introduces sufficient variability into the dataset, capturing diverse perspectives of Bahnar language. This midpoint selection avoids the extremes of overly aggressive or overly conservative augmentation, ensuring a reasonable balance between enriching the dataset with diverse instances and maintaining a degree of stability essential for effective model training.

It is worth noting that the experiments were conducted with various alpha values spanning the range from 0.0 to 1.0. However, a value of 0.5 showcases its suitability for maintaining a balance between augmentation-induced diversity and dataset stability. Additionally, due to limited resources for both training and augmenting data, adopting a cautious and proactive approach, an alpha value of 0.5 is chosen as a safe and resource-efficient strategy.

The bilingual dictionary houses $13,029$ words aligning both the Bahnar language and Vietnamese dictionaries. After augmenting the original dataset, the expanded corpus undergoes a thoughtful division. This refined dataset undergoes strategic partitioning, allocating 80% for training purposes. Within this training phase, sentence chunking is activated and executed. The remaining 20% of the dataset is intended for training with complete sentences, ensuring a comprehensive approach that embraces both segmented and intact linguistic contexts. Overall, this division is focused on enhancing the model's adaptability and proficiency across diverse linguistic structures.

Baselines To assess the performance of our model, we conducted 6 experiments, wherein five baseline models were employed for comparison against our proposed model. Our baselines include:

- **Transformer:** We reproduce a full stack of 6 encoder and decoder layers in the vanilla Transformer (Vaswani et al. 2017)
- **PhoBERT-fused NMT:** A BERT-fused NMT model with the replacement of PhoBERT

Model	BLEU
Transformer	28.78
PhoBERT-fused NMT	38.49
BARTphoEncoderPGN	47.91
Loanformer	42.63
PE-PD-PGN	49.00
BN-BARTpho	**49.20**

Table 1: BLEU Scores for each model

- **Loanformer:** A model that combines PhoBERT-fused NMT and the masked Pointer Generator Network. This is a model that we are also developing in our team
- **BARTphoEncoderPGN:** Similar to Loanformer, but PhoBERT is replaced by the encoder of BARTpho.
- **PE-PD-PGN:** A PE-PD-fused NMT model combined with the masked Pointer Generator Network

Setting In the training phase, we maintain a dropout rate of 0.1 and aggregate gradients over 64 batches, each comprising 4 instances, per weight update. To establish a sensible boundary, we cap the maximum length of both source and target sequences at 256, and introduce label smoothing with ϵ_{ls} set at 0.1. Our choice of optimization tools, including the Adam optimizer and learning rate scheduler, adheres to the conventions outlined in (Zhu et al. 2020). All of the models are trained through 21 epochs, involving approximately $4,000$ weight updates for each model.

Result

After conducting experiments, the authors computed the corresponding BLEU scores for each model, as shown in Table 1. It is evident from the results that the BN-BARTpho model achieved the highest BLEU score, indicating its efficiency in our considering methodology.

These results highlight the decent performance of the BN-BARTpho model in our research endeavors. Since Bahnar language is a low-resource language, it has a limited vocabulary and few synonyms. This means that when translating from a high-resource language like Vietnamese to Bahnar language, there are fewer variations in the possible translations for each n-gram. As a result, it is easier for the translation model to generate accurate translations and achieve a high BLEU score.

Besides, researchers has conducted experiments for the translation from Vietnamese to Bahnar to evaluate the effect of using each of the MTL DA auxiliary tasks, with the combination of the best-performing ones and the sentence boundary approach. Table 2 reports the translation performance, measured in terms of BLEU score prediction.

First, the baseline is the evaluation results when training and testing without any augmentation method are applied. The results show that the MTL DA approach consistently outperforms the baseline system except for method *source*. In general, the auxiliary tasks *swap*, *token*, and *replace* are the best-performing ones. *reverse* may give a lower performance result than these three methods above, which suggests that abnormal word order could negatively influence the main task. In contrast, *source* has its worst performance, which indicates that the translation task could be affected by introducing a completely different vocabulary in the target.

Method	BLEU
baseline	49.20
swap	**52.94**
token	**51.80**
source	2.59
reverse	39.79
replace	51.33
swap+replace	51.57
swap+token	**53.91**
replace+token+swap	52.89
sentence boundary	**54.31**

Table 2: BLEU scores obtained with the baseline; MTL DA approach, using different auxiliary tasks and combinations of them; and sentence boundary augmentation

Interestingly, using each two of the three best auxiliary tasks together further improves the performance, achieving well-performed results with higher BLEU scores of 51.57 (swap+replace) and 53.91 (swap+token) points than the baseline. The combination of **swap+token** gives the best performance. Although the researcher has combined all three best methods, the results still cannot outperform the combination of token and swap. Overall, all combinations have improved the BLEU score; this implies that various auxiliary tasks impact the encoder in distinct manners and exhibit a sense of complementarity.

Considering the same training configuration, the result of *sentence boundary augmentation* can be paired with any auxiliary task from MTL DA and their combinations. It indicates that *sentence boundary augmentation* can perform well in the context of low-resource translation. It can also utilize the available limited resources. For low-resource machine translation augmentation, the nosing-based method applied at the phrase or sentence level has demonstrated superior performance compared to the nosing-based method applied at the word level.

Conclusion

Our study emphasizes the utmost significance of revitalizing Bahnar language, a cultural and historical treasure of Vietnam's Bahnar people. We propose a NMT model for Vietnamese-Bahnar, paving the way for inter-generational connection and cross-cultural communication. However, translating Vietnamese into Bahnar language remains as a serious challenges: Scarce data, limited vocabulary, and lack of fine-tuning resources. Our multifaceted approach tackles these hurdles by leveraging transfer learning from pre-trained models, capitalizing on the linguistic kinship between Vietnamese and Bahnar. We also implement tailored augmentation strategies to fine-tune the NMT system for the nuances of Vietnamese-Bahnar. This approach demonstrably outperforms baseline models, not only advancing translation quality but also showcasing NMT's potential for revitalizing low-resource languages.

Acknowledgements

This research is funded by Ministry of Science and Technology (MOST) within the framework of the Program ”Supporting research, development and technology application of Industry 4.0” KC-4.0/19-25 – Project “Development of a Vietnamese- Bahnaric machine translation and Bahnaric text-to-speech system (all dialects)” - KC-4.0-29/19-25. We are also grateful to Mr. Nguyen Tran Quy and his linguistics research group for the Vietnamese - Bahnaric corpus collected during their field research trips.

References

Arora, S.; Liang, Y.; and Ma, T. 2017. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In *International Conference on Learning Representations*.

Artetxe, M.; Labaka, G.; Agirre, E.; and Cho, K. 2017. Unsupervised Neural Machine Translation. *CoRR*, abs/1710.11041.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Caswell, I.; Chelba, C.; and Grangier, D. 2019. Tagged Back-Translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, 53–63. Florence, Italy: Association for Computational Linguistics.

Chapelle, O.; Scholkopf, B.; and Zien, A., Eds. 2009. Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006) [Book reviews]. *IEEE Transactions on Neural Networks*, 20(3): 542–542.

Currey, A.; Miceli Barone, A. V.; and Heafield, K. 2017. Copied Monolingual Data Improves Low-Resource Neural Machine Translation. In *Proceedings of the Second Conference on Machine Translation*, 148–156. Copenhagen, Denmark: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

Edunov, S.; Ott, M.; Auli, M.; and Grangier, D. 2018. Understanding Back-Translation at Scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 489–500. Brussels, Belgium: Association for Computational Linguistics.

Fadaee, M.; Bisazza, A.; and Monz, C. 2017. Data Augmentation for Low-Resource Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 567–573. Vancouver, Canada: Association for Computational Linguistics.

Gal, Y.; and Ghahramani, Z. 2016. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, 1027–1035. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510838819.

Gao, F.; Zhu, J.; Wu, L.; Xia, Y.; Qin, T.; Cheng, X.; Zhou, W.; and Liu, T.-Y. 2019. Soft Contextual Data Augmentation for Neural Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5539–5544. Florence, Italy: Association for Computational Linguistics.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. In *Neural Computation, 9(8):1735–1780*.

Lample, G.; Ott, M.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2018. Phrase-Based & Neural Unsupervised Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 5039–5049. Brussels, Belgium: Association for Computational Linguistics.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 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*, 7871–7880. Online: Association for Computational Linguistics.

Li, D.; I, T.; Arivazhagan, N.; Cherry, C.; and Padfield, D. 2021. Sentence Boundary Augmentation for Neural Machine Translation Robustness. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7553–7557.

Liu, R.; Xu, G.; Jia, C.; Ma, W.; Wang, L.; and Vosoughi, S. 2020. Data Boost: Text Data Augmentation Through Reinforcement Learning Guided Conditional Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9031–9041. Online: Association for Computational Linguistics.

Luong, T.; Pham, H.; and Manning, C. D. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1412–1421. Lisbon, Portugal: Association for Computational Linguistics.

Luque, F. M. 2019. Atalaya at TASS 2019: Data Augmentation and Robust Embeddings for Sentiment Analysis. In *IberLEF@SEPLN*.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. In Bengio, Y.; and LeCun, Y., eds., *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Nguyen, D. Q.; and Nguyen, A. T. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1037–1042.

Nguyen, D. Q.; Nguyen, D. Q.; Vu, T.; Dras, M.; and Johnson, M. 2018. A Fast and Accurate Vietnamese Word Segmenter. In Calzolari, N.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Hasida, K.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; Piperidis, S.; and Tokunaga, T., eds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P.; Charniak, E.; and Lin, D., eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21(1).

Sánchez-Cartagena, V. M.; Esplà-Gomis, M.; Pérez-Ortiz, J. A.; and Sánchez-Martínez, F. 2021. Rethinking Data Augmentation for Low-Resource Neural Machine Translation: A Multi-Task Learning Approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 8502—8516. Association for Computational Linguistics.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, 3104–3112. Cambridge, MA, USA: MIT Press.

Takase, S.; and Kiyono, S. 2021. Rethinking Perturbations in Encoder-Decoders for Fast Training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5767–5780. Online: Association for Computational Linguistics.

Takase, S.; and Kiyono, S. 2023. Lessons on Parameter Sharing across Layers in Transformers. In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, 78–90. Toronto, Canada (Hybrid): Association for Computational Linguistics.

Tan, Z.; Wang, S.; Yang, Z.; Chen, G.; Huang, X.; Sun, M.; and Liu, Y. 2020. Neural machine translation: A review of methods, resources, and tools. *AI Open*, 1: 5–21.

Tran, N. L.; Le, D. M.; and Nguyen, D. Q. 2022. BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association*.

Vaswani, A.; Shazeer, N. M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NIPS*.

Vázquez, R.; Raganato, A.; Tiedemann, J.; and Creutz, M. 2019. Multilingual NMT with a Language-Independent Attention Bridge. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, 33–39. Florence, Italy: Association for Computational Linguistics.

Vu, T.; Nguyen, D. Q.; Nguyen, D. Q.; Dras, M.; and Johnson, M. 2018. VnCoreNLP: A Vietnamese Natural Language Processing Toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 56–60. New Orleans, Louisiana: Association for Computational Linguistics.

Wei, J.; and Zou, K. 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)*, 6382–6388. Hong Kong, China: Association for Computational Linguistics.

Xia, Y.; Qin, T.; Chen, W.; Bian, J.; Yu, N.; and Liu, T.-Y. 2017. Dual Supervised Learning. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 3789–3798. PMLR.

Xie, Z.; Wang, S. I.; Li, J.; Lévy, D.; Nie, A.; Jurafsky, D.; and Ng, A. Y. 2017. Data Noising as Smoothing in Neural Network Language Models. In *International Conference on Learning Representations*.

Zhang, H.; Qiu, S.; Duan, X.; and Zhang, M. 2020. Token Drop mechanism for Neural Machine Translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4298–4303. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Zhang, X.; Zhao, J. J.; and LeCun, Y. 2015. Character-level Convolutional Networks for Text Classification. In *NIPS*.

Zhang, Z.; Wu, S.; Liu, S.; Li, M.; Zhou, M.; and Chen, E. 2019. Regularizing Neural Machine Translation by Target-Bidirectional Agreement. In *AAAI Conference on Artificial Intelligence*.

Zhu, J.; Xia, Y.; Wu, L.; He, D.; Qin, T.; Zhou, W.; Li, H.; and Liu, T. 2020. Incorporating BERT into Neural Machine Translation. In *International Conference on Learning Representations*.