Low-Resource NLP: Building an English-Bahnar Machine Translation Model

Group Members:
Phan Thi Hien Chi, 21chi.pth@vinuni.edu.vn
Ha Phuong Thao, 20thao.hp@vinuni.edu.vn
Nguyen Thai Uyen, 21uyen.nt@vinuni.edu.vn

1 Introduction

The Bahnar language, an Austroasiatic language spoken predominantly in the Central Highlands of Vietnam, with approximately 140,000 speakers, is not merely a means of communication but a vessel for the community's traditions and history [7]. However, in the digital era, low-resource languages like Bahnar face the threat of digital extinction due to their scant presence in the rapidly evolving technological landscape.

Machine translation systems have the potential to bridge this gap, yet developing such systems for these languages is fraught with challenges. The scarcity of bilingual corpora, the complex linguistic features inherent to Bahnar, and the lack of previous research on computational models for the language exacerbate the difficulty of this task.

This project aims to tackle the aforementioned challenges by developing a machine translation system for the Bahnar language. The objective is to provide a tool that facilitates communication and access to information for Bahnar speakers, and to contribute to the field of computational linguistics with methodologies gained from working with a low-resource language.

The code for our project is available at https://github.com/chiphan1110/Bahnar-MachineTranslation.git

2 Our Approaches

2.1 Dataset Construction

2.1.1 Data Collection

The foundation of any neural machine translation system is a robust, bilingual dataset. For low-resource languages such as Bahnar, acquiring such data presents unique challenges due to limited availability. To address this, we implemented a multifaceted data collection strategy:

- **Bahnar-English Dataset:** Our primary source was an online bilingual Bible, where we systematically crawled English-Bahnar parallel sentences from YouVersion [3]. It is the only substantial corpus available in both Bahnar and English.
- **Bahnar-Vietnamese Dataset:** Given the regional proximity and availability, we expanded our data collection to include Bahnar-Vietnamese pairs through the following methods:
  - **Bible:** The Bahnar-Vietnamese bible version was also crawled from YouVersion.
  - **Educational Resources**: We manually processed OCR on Bahnar textbooks, due to the lack of tools supporting Bahnar language. The textbooks were retrieved from SIL International, consisting of Bahnar Language Lessons [1] and Bahnar First Grade Primer [2]
  - **Legal Documents:** We extracted bilingual text from legal documents available in PDF format [4]. The law of the Bahnar people also contains a variety of contexts and subdomains (like criminal law, civil rights, etc.), offering a diverse training set.
  - **Online News:** We leveraged a GitHub repository from [9] that aggregates news articles in Bahnar, providing us with contemporary and topical language usage.

2.1.2 Data Cleaning and Pre-processing

The raw data collected from various sources required meticulous cleaning and pre-processing to ensure its quality and usability for machine translation. The following steps were taken:

- **Web-Crawled Data Pre-processing**: We used an available tool [5] to scrape the bible verses directly from the website and made our own modifications in the code from this Github repository [16] to pair verse by verse and sentence by sentence (parallel corpus) for later use in the training phase.

	News	Education	Laws	Bible	Total pairs
Ba-En original	N/A	N/A	N/A	22224	22224
Ba-Vi	11839	739	5326	17071	34,975
Ba-En augmented	11839	739	5326	17071	34,975

Table 1: Dataset statistics

- **PDF Data Extraction**: Textual content was extracted from PDF documents using the PyMuPDF library. We implemented routines to systematically remove headers, footers, and any non-textual elements that do not contribute to the translation task. Special attention was given to the normalization of Bahnar language characters that have specific diacritical marks. We standardized these across our dataset to avoid any discrepancies that could affect the translation quality.
- **Sentence Alignment**: The cleaned English, Bahnar, and Vietnamese texts were then subjected to sentence alignment. This process pairs sentences from the source language with their equivalent translations in the target language, a critical step for any parallel corpus.
- **Data Filtering and Finalization**: The culminating pre-processing phase involved curating the dataset to balance the distribution of sentence lengths and complexity. This was to ensure that the model would not be biased toward either simple or complex sentence structures.

The final dataset includes 22,424 Bahnar-English sentence pairs and 34,975 Bahnar-Vietnamese which then we separate 2000 pairs for testing and the remaining part for training and validation in each dataset. The detailed dataset statistics are reported in the Table 1.

2.1.3 Data Augmentation via Vietnamese pivot language

Recognizing the limited volume of direct English-Bahnar parallel texts, we applied a data augmentation approach using Vietnamese as a pivot language. We first translated our English corpus into Vietnamese using a high-accuracy machine translation API provided by VinAI [8]. This translation served as an intermediate step as the newly created English-Bahnar parallel sentences were added to the existing corpus, carefully curated to enrich linguistic diversity and complexity. To assess the impact of data augmentation, a control group of 2000 sentence pairs from the original dataset was maintained for testing. Post-augmentation, our dataset exhibited substantial growth in size. The augmented dataset's statistics are detailed in Table 1.

2.2 Machine Translation Model

2.2.1 Problem Formulation:

A neural machine translation system is a neural network that directly models the conditional probability $p(\mathbf{y}|\mathbf{x})$ of translating source sentence $\mathbf{x} = x_1, \cdots, x_n$ to a target sentence $\mathbf{y} = y_1, \cdots, y_m$
Formally, the objective of our NMT system can be defined as:

$$P(\mathbf{y}|\mathbf{x};\Theta) = \prod_{j=1}^{m} P(y_j|\mathbf{y}_{<j},\mathbf{x};\Theta) \tag{1}$$

Where:

- $\mathbf{x} = (x_1, ..., x_n)$ is the input sentence in the source language with length n
- $\mathbf{y} = (y_1, ..., y_m)$ is the output sentence in the target language with length m.
- $\mathbf{y}_{<j}$ denotes all the tokens before position j in the target sentence.
- Θ represents the parameters of the Neural Machine Translation model.

The model parameters Θ are learned by minimizing the negative log-likelihood of the bilingual dataset $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{N}$, where N is the number of sentence pairs in the dataset:

$$\mathcal{L}(\Theta) = -\sum_{i=1}^{N} \log P(\mathbf{y}^{(i)}|\mathbf{x}^{(i)};\Theta) \tag{2}$$

2.2.2 Transformer-based Machine Translation model:

Our machine translation system utilizes the Transformer [17] model , leveraging its proven efficacy in neural machine translation. The Transformer distinguishes itself with an architecture comprising stacked encoder and decoder layers, which facilitate parallel processing and capture complex dependencies in the data. Each encoder consists of a multi-head self-attention mechanism and a feed-forward network, enhanced with layer normalization and residual connections for stability and training efficiency. The decoder extends this design with an additional attention layer that integrates the encoder's output. We opted for the Transformer model due to its state-of-the-art performance and the flexibility of its architecture, which allows for effective learning of linguistic representations and translation mappings. Training was executed using the OpenNMT [6] toolkit, chosen for its robust support of sequence learning tasks and its widespread adoption in the NMT community.

2.2.3 Helsinki and mBART-50 Pre-trained Models

In addition to our custom Transformer models, we incorporated Helsinki-NLP's [15] pretrained models for Bahnar-English pairs and facebook/mbart-large-50 [14] pretrained models for Vietnamese-English pairs in our machine translation system. Inspired by the Transfer Learning approach from Zoph et al. (2016) [19], a "parent" model was trained using a language pair with more abundant resources, its weights were then adapted to create a "child" model, which further refined using our low-resource language pairs (Bahnar-English and Bahnar-Vietnamese). As similar parent and child languages yielded better results, we used Vietnamese-English and Khmer-Vietnamese as the "parent" models for Bahnar-English and Bahnar-Vietnamese models, respectively. We expected some similarities between Bahnar, Khmer, and Vietnamese because they all belong to the Austroasiatic language family. By fine-tuning the Helsinki-NLP and Facebook's mBART-50 models on our Bahnar language dataset, we aimed to leverage the benefits from the cross-linguistic patterns learned during the pretraining on high-resource languages to enhance our translation models.

3 Experiments

3.1 Experiment Setup

Prior to training our models, we employed Byte Pair Encoding (BPE) subword tokenization to address the open vocabulary problem in neural machine translation [12]. BPE allows us to efficiently represent rare and unknown words as sequences of subword units. This step is crucial for achieving better translation quality, especially in the context of low-resource languages where the dataset may not cover an extensive vocabulary.

We adhered to a uniform training protocol for all our supervised Neural Machine Translation (NMT) models. Training was executed utilizing the Adam optimizer with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.998$, and $\epsilon = 10^{-9}$, following the approach described in the Transformer paper [17]. Our learning rate was initially set to 1.7×10^{-7}, undergoing a warmup over the first 1000 steps, and subsequently transitioning to an inverse square root decay schedule.

In our model architecture, which is based on the Transformer, we incorporated the Rectified Linear Unit (ReLU) as the activation function for both the feed-forward network and the self-attention mechanism within the encoder-decoder layers. We applied a consistent dropout rate of 0.1 across all layers to mitigate overfitting. Model training was conducted with batches containing 4096 tokens. Early stopping was set to terminate training if no improvement was detected after 4 validation checks. Our models were trained for a total of 3000 steps, with validations performed every 1000 steps to monitor and evaluate model performance.

For the Helsinki-NLP and mBART-50 pretrained models, we loaded the models and tokenizers using HuggingFace's transformer library [18] and incorporated a consistent weight decay of 0.015 for the optimizer to avoid overfitting. Models training was conducted with batches containing 32 tokens for Bahnar-English pairs and 2 tokens for Bahnar-Vietnamese pairs with validations performed at every step. For Bahnar-English pairs using the Helsinki-NLP's models the learning rate is 2.0×10^{-5} with 5 steps for original data and 10 steps for augmented data. For Bahnar-Vietnamese pair using mBART-50, the learning rate is 4.0×10^{-5} with 2 training steps.

3.2 Evaluation Metrics

To rigorously assess the performance of our machine translation system, we employed different automatic evaluation metrics: SacreBLEU [11], chrF [10] and TER [13]. SacreBLEU provides a consistent and comparable BLEU score by standardizing the calculation parameters. chrF is utilized for its ability to compute character n-gram F-scores, offering insights into morphological richness. Translation Edit Rate (TER) measures the edit distance to convert a translated output into a reference translation, with lower scores indicating higher quality. Each of these metrics is language-independent and freely available, enabling a comprehensive and fair evaluation of machine translation outputs. The

Model	Ba-to-Vi			Vi-to-Ba		
	BLEU↑	**CHRF↑**	**TER↓**	**BLEU↑**	**CHRF↑**	**TER↓**
mBART-50	**49.61**	**59.51**	**34.78**	**39.08**	**55.73**	**54.49**
Transformer-based	22.09	41.76	48.49	21.93	47.58	57.97

Table 2: Experimental results of translation models for Bahnar-Vietnamese dataset

Dataset	**Model**	**Ba-to-En**			**En-to-Ba**		
		BLEU↑	**CHRF↑**	**TER↓**	**BLEU↑**	**CHRF↑**	**TER↓**
Original Ba-En dataset	Helsinki-NLP	40.20	52.74	42.20	24.43	50.21	64.40
	Transformer-based	18.61	39.00	79.08	16.50	39.06	76.57
Augmented Ba-En	**Helsinki-NLP**	**43.73**	**55.62**	**41.16**	**38.42**	**51.10**	**62.51**
	Transformer-based	20.53	25.27	71.65	19.73	41.05	71.79

Table 3: Experimental results of translation models for the original & augmented Bahnar-English dataset

combination of these metrics allows us to capture various aspects of translation quality, including adequacy, fluency, and lexical choice.

4 Results and Discussion

4.1 Automatic Evaluation

For the Bahnar-to-Vietnamese translation task, the mBART-50 pre-trained model outperformed the Transformer-based model with a BLEU score of 49.61 compared to 22.09, and other metrics also demonstrate similar results. The pre-trained model also performs better in the Vietnamese-to-Bahnar translation task with a BLEU score of 39.08, chrF of 55.73, and TER of 54.49. The full experimental results are shown at Table 2. For translations involving English, as detailed in Table 3, the Helsinki-NLP pre-trained models also surpassed the Transformer-based models on both original and augmented Bahnar-English datasets.

These results demonstrate the effectiveness of pre-trained models on machine translation tasks. The Helsinki-NLP and mBART-50 pre-trained models consistently outperformed the Transformer-based models across all language pairs. This performance discrepancy can be attributed to the extensive multilingual training data that pretrained models have been exposed to, allowing them to generalize better even when applied to a low-resource language like Bahnar.

The augmented dataset, which used Vietnamese as a pivot language, yielded significant improvements in translation quality for all models, demonstrating the potential of this method to enhance the quality of machine translations in low-resource language scenarios. This improvement highlights the significant role data augmentation plays in machine translation, particularly when addressing the challenges of under-resourced languages. However, the results also highlight the challenges faced by the Transformer-based models trained from scratch. Without the benefit of large-scale, diverse pretraining, these models may struggle to achieve high translation quality, underscoring the importance of pretraining and data augmentation in the context of low-resource languages.

4.2 Human Qualitative Evaluation

To see how the model can capture the nuance and context of the original version, we pick 5 random pairs in the evaluation set for the transformer model and 3 pairs for the pretrained model. The transformer model with augmentation data provides a more interpretive translation for modern comprehension, as seen in "They looked, but there was none to save;" versus the "ba-en" version's literal "There they cry, but none giveth answer. Meanwhile, the model without augmentation data retains the source text's structure and wording, favoring a 'word-for-word' approach that may serve academic or literary uses better. The choice between models depends on the need for clarity or fidelity to the original text. For pretrained model, the results suggested that augmented data did improve the translation quality.

Original Translation	Ba-En with augmented data	Ba-En
They cried, but there was none to save them:	They looked, but there was none to save;	They cry, but none giveth answer,
The heavens declare the glory of God;	The heaven and heaven shall declare his glory:	The heaven and the heavens shall be praised:
The clouds also dropped water	Clouds and hail are round about	Or whether it be rain
These are the Gods that smote the Egyptians with all the plagues in the wilderness	This is the gods of the Egyptians, which were in the wilderness	These are the countries which the LORD smote the Egyptians in the wilderness
They left off speaking	But they are utterly consumed with words	All his words are no more

Table 4: Comparison of Transformer model

Original Translation	Ba-En with augmented data	Ba-En
And over the host of the tribe of the children of Simeon was Shelumiel the son of Zurishaddai	And over the host of the children of Simeon was Shelumiel the son of Zuisadai	And over the sheep of Simeon, Shelumiel the son of Zurishaddai
Josiah was eight years old when he began to reign, and he reigned in Jerusalem one and thirty years	And Josiah was eight years old when he began to reign, and he reigned thirty years in Jerusalem	And Josiah was afraid, that he should be strong, and that he might be in Jerusalem one and another
And we said unto him, We are true men; we are no spies	And we said unto him, We are good men, and will not look upon thy land	And we are sick; we are full

Table 5: Comparison of Helsinki-NLP pretrained model

5 Conclusion

This study presented a comprehensive analysis of machine translation systems for the Bahnar language, demonstrating the effectiveness of pre-trained models and the significance of data augmentation in enhancing translation quality. While the Helsinki-NLP models showed promising results, there remains a gap in the performance of Transformer-based models trained from scratch. The challenges are particularly pronounced when dealing with a low-resource language, where data scarcity and quality are pivotal concerns. Our work still has limitations. The data sources we have constructed, while useful, may not fully represent the diversity of contexts in which Bahnar is used. Texts from religious scriptures or legal documents may not include colloquial or modern usage which is vital for a comprehensive language model. The evaluation metrics employed, while standard, do not always correspond to human judgments of translation quality. Our qualitative evaluations are limited and based on our subjective observations due to time constraints. To overcome these limitations, we plan to explore more data augmentation methods such as backtranslation and try more pretrained models in the future.

6 Contribution an Acknowledgement

Team member contribution:

- Phan Thi Hien Chi: prepare and preprocess data (legal documents), build the Transformer model for Bahnar-English.
- Ha Phuong Thao: prepare and preprocess data (education documents and news), build the models for Bahnar-Vietnamese.
- Nguyen Thai Uyen: prepare and preprocess data (bible), fine-tune the pre-trained models for Bahnar-English and Bahnar-Vietnamese.

We would like to express our gratitude to Prof. David Harrison for providing the Bahnar-English dictionary during the project's early stages, as well as the COMP3020 teaching team for their helpful support and guidance throughout our projects.

References

[1] Bahnar language lessons pleiku.

[2] Bahnar teachers guide for primer 1.

[3] The bible app | bible.com.

[4] Luat tuc bahnar - ttn.

[5] IonicaBizau. Ionicabizau/bible-scraper: retrieve verses from bible.com/youversion.

[6] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. Opennmt: Open-source toolkit for neural machine translation. In *Proc. ACL*, 2017.

[7] Dao Bui Minh. *The Bahnar people in Vietnam*. World Publishers, 2011.

[8] Thien Hai Nguyen, Tuan-Duy H. Nguyen, Duy Phung, Duy Tran-Cong Nguyen, Hieu Minh Tran, Manh Luong, Tin Duy Vo, Hung Hai Bui, Dinh Phung, and Dat Quoc Nguyen. A Vietnamese-English Neural Machine Translation System. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association: Show and Tell (INTERSPEECH)*, 2022.

[9] Nhatkhangcs. Nhatkhangcs/eaai24-official-bahnaric-dataset: This is the official binh dinh bahnaric dataset, used for training mentioned in our paper work.

[10] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[11] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[12] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units, 2016.

[13] Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August 8-12 2006. Association for Machine Translation in the Americas.

[14] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. 2020.

[15] Jörg Tiedemann and Santhosh Thottingal. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal, 2020.

[16] TraMiu. Tramiu/bible-datascraper: retrieve bahnar and english verses from bible.com/youversion with modifications.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[18] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[19] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. *CoRR*, abs/1604.02201, 2016.