Science & Technology Development Journal – Engineering and Technology 2023, 6(4):2099-2117

Open Access Full Text Article

Research article

Leveraging sentence-oriented augmentation and transformer-based architecture for Vietnamese-Bahnaric translation

Sang T. Nguyen[1,2], Nguyen Q. Pham[1,2,*], Tho Quan[1,2]

[1]Falcuty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet, District 10, Ho Chi Minh City, Vietnam

[2]Vietnam National University Ho Chi Minh City (VNU-HCM), Linh Trung Ward, Thu Duc City, Ho Chi Minh City, Vietnam

Correspondence

Nguyen Q. Pham, Falcuty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet, District 10, Ho Chi Minh City, Vietnam

Vietnam National University Ho Chi Minh City (VNU-HCM), Linh Trung Ward, Thu Duc City, Ho Chi Minh City, Vietnam

Email: quocnguyenkh@gmail.com

History
- Received: 15-10-2023
- Accepted: 25-3-2024
- Published Online: 13-5-2024

DOI : https://doi.org/10.32508/stdjet.v6i4.1284

Check for updates

ABSTRACT

The Bahnar people, an ethnic minority in Vietnam with a rich ancestral heritage, possess a language of immense cultural and historical significance. The government places a strong emphasis on preserving and promoting the Bahnaric language by making it accessible online and encouraging communication across generations. Recent advancements in artificial intelligence, such as Neural Machine Translation (NMT), have brought about a transformation in translation by improving accuracy and fluency. This, in turn, contributes to the revival of the language through educational efforts, communication, and documentation. Specifically, NMT is pivotal in enhancing accessibility for Bahnaric speakers, making information and content more readily available. Nevertheless, the translation of Vietnamese into Bahnaric faces practical challenges due to resource constraints, especially given the limited resources available for the Bahnaric language. To address this, we employ state-of-the-art techniques in NMT along with two augmentation strategies for domain-specific Vietnamese-Bahnaric translation task. In a multi-task data augmentation approach, new sentence pairs are generated through transformations. These augmented sentences are employed as auxiliary tasks within a multi-task framework during training. The objective is to introduce fresh contexts where the target prefix alone does not provide sufficient information for predicting the next word accurately. His approach enhances the encoder's capabilities and compels the decoder to focus more on the source representations from the encoder. On the other hand, the sentence boundary augmentation method extends the application of the noising-based approach beyond the word level to include sentence-level augmentation. In neural machine translation, handling errors related to grammatical structure and sentence boundaries poses significant challenges to ensure robustness. Through thoroughly examining errors, it becomes evident that sentence boundary segmentation has the most substantial impact on translation quality. To enhance segmentation robustness, a straightforward data augmentation strategy is devised. Importantly, both approaches are flexible and can be used with various neural machine translation models. Additionally, they do not require complex data preprocessing steps, the training of additional systems, or the acquisition of extra data beyond the existing training parallel corpora.

Key words: Data augmentation, low-resource neural machine translation, machine translation, Bahnar language

INTRODUCTION

The Bahnar people, sometimes called Ba-Na, represent a distinct ethnic minority in Vietnam's diverse spectrum of ethnic groups. At present, the Vietnamese government is actively striving to enable their full participation in the wider society. This effort primarily concentrates on enhancing their participation in social and cultural aspects, as well as in the fields of education and science. As a component of this program, there is a significant focus on translating crucial documents into the Bahnar language.

Therefore, machine translation has been addressed as a possible solution for translating Vietnamese to Bahnar. With the emergence of deep learning in recent years, Neural Machine Translation [1,2] has become a new model and become the mainstream of machine translation. NMT helps establish an accessible means of communication for Bahnar people. This research is not merely focused on scientific techniques; it is also a means of preserving the national language and paying tribute to the spiritual values and culture of the Bahnar people.

While the availability of large parallel corpora significantly impacts how an NMT system performs, the Bahnar language itself is a low-resource language [3], which can make the system suffer from poor translation quality [4]. Therefore, Data Augmentation (DA) [5] needs to be involved in the project to generate extra data points from the empirically observed train-

Cite this article : Nguyen S T, Pham N Q, Quan T. **Leveraging sentence-oriented augmentation and transformer-based architecture for Vietnamese-Bahnaric translation**. *Sci. Tech. Dev. J. – Engineering and Technology* 2023, 6(4):2099-2117.

ing set to train the NMT model. With DA, the performance of the translation system can be significantly improved, all while making efficient use of the current resources. Data augmentation was first widely applied in the computer vision field and then used in natural language processing, achieving improvements in many tasks. DA helps to improve the diversity of training data, thereby helping the model anticipate the unseen factors in testing data. DA applications in natural language processing have been investigated in recent years, and the most well-known fields are text classification [6–8], text generation (including NMT) [9–11], and structure prediction [12,13]. DA is still a super common and all-over-the-place approach in NMT, which samples some fake data distribution $P_f(X')$ using some common methods (Figure 1) based on real data distribution $P_r(X)$, where X_f^1, X_f^2 refer to augmented data generated from real data using common approaches, such as replacing, swapping.

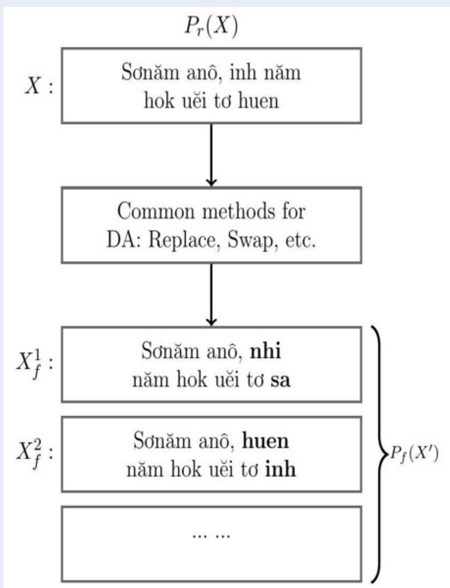

Figure 1: The commonly used methods of DA for NMT

In this paper, researchers have explored and tested two DA methods. Both of these approaches can produce parallel sentences that, even though they are highly improbable based on the data distribution, consistently enhance the quality of the resulting NMT system. The first method follows the main idea of the method Easy Data Augmentation (EDA) [7], which is a framework that contains a simple set of chosen DA techniques for natural language processing. However, the DA procedures employed in this method are inspired by the findings of multi-task learning data augmentation (MTL DA) research [14]. As a result, researchers have chosen to name it MTL DA as well. This approach generates artificial target sentences with the purpose of reinforcing the encoder. The second approach, drawing inspiration from the work of [15], takes a distinct path by extending the application of the noise-based augmentation technique beyond the word level. Researchers have examined the potential of this method in different contexts beyond the original research, which primarily focused on low-resource NMT. Several extensive experiments have been conducted and delved deeply into the hyperparameters that govern the behavior of this method.

Neither method necessitates complex preprocessed procedures, training additional systems, or utilizing data beyond the existing training parallel corpora. Experiments were conducted to assess the performance enhancement achieved by these methods compared to the previous research (EDA) and the widely recognized augmentation technique (word replacement using semantic embedding). In addition to evaluating their effectiveness on the entire dataset, both methods have been specifically researched and employed to address certain sentence forms that pose significant challenges in the translation task.

Contributions of our work are as follows:

- Studying and implementing two data augmentation techniques, including multitask learning data augmentation and sentence boundary augmentation, to enhance low-resource language data, thus improving the performance of the NMT model.
- A more in-depth study was conducted on each data augmentation method by selecting the most suitable data augmentation operations and parameters for the Vietnamese to Bahnar translation task.
- Effectively applying the mentioned data augmentation methods helps in improving common errors (wrong collocation and word-by-word) in Vietnamese-Bahnar translation.

The remainder of the paper is organized as follows. The next section briefly provides background knowledge of neural machine translation, data augmentation, and its primary techniques. After that, Section "Bhanar language" gives an overview of the Bhanar language and the differences between Vietnamese and Bahnar. Section "Related works" presents a review of the most relevant works in the area of DA for NMT. The details of two methods MTL DA and sentence boundary augmentation, are shown in Section

"Approaches", whereas Section "Experiments and discussion" describes the experimental settings and discusses the results obtained. Finally, the paper ends with some concluding remarks in Section "Conclusion".

PRELIMINARIES

Neural machine translation

Overview

Machine Translation (MT) [16] is a major sub-field of Natural Language Processing (NLP) [17] that focuses on translating human languages automatically by using a computer. In the early stages, machine translation relies heavily on manual translation rules and linguistic knowledge. However, because the nature of language is significantly complicated, it is impossible to cover all irregular cases with just hand-crafted translation rules. During the development process of MT, more and more large-scale parallel corpora appeared. With the data-driven approaches, Statistical Machine Translation (SMT) [18] has replaced the original rule-based translation due to its availability to study latent factors such as word alignment or phrases directly from corpora. However, SMT is still far from expectations because it is unable to model long-distance word dependencies. With the emergence of deep learning, Neural Machine Translation has become a new model and replaced SMT to become the mainstream of MT.

Modeling

Assuming a source sentence $x = \{x_1, \ldots, x_S\}$ and a target sentence $y = \{y_1, \ldots, y_S\}$ are given. By using the chain rule, the conditional distribution of a standard NMT [19,20] can factorize a sentence-level translation probability as a product of word-level probabilities from left-to-right (L2R) as:

$$P(y|x) = \prod_{t=1}^{T} P(y_t|y_0, \ldots y_{t-1}, x) \quad (1)$$

NMT models that conform to Eq. 1 are referred to as L2R autoregressive NMT [1, 2] for the prediction at time-step t is taken as input at time-step t+1.

NMT normally uses maximum log-likelihood (MLE) as the training objective function, which is usually used for estimating the parameters of a probability distribution. Given the training corpus $D = \left\{\left(x^{(s)}, y^{(s)}\right)\right\}_{s=1}^{S}$, the goal of training is to find a set of model parameters that maximize the log-likelihood on the training set:

$$\widehat{\theta}_{MLE} = \underset{x}{argmax}\left(L(\theta)\right), \quad (2)$$

where the log-likelihood is defined as

$$L(\theta) = \sum_{s=1} \log P\left(y^{(s)}|x^{(s)};\theta\right) \quad (3)$$

By the back-propagation algorithm, the gradient of L can be computed with respect to θ. NMT model training usually adopts the stochastic gradient search (SGD) algorithm. Instead of computing gradients on the full training set, SGD computes the loss function and gradients on a mini-batch of the training set. The plain SGD optimizer updates the parameters of an NMT model with the following rule:

$$\theta \leftarrow \theta - \alpha \nabla L(\theta) \quad (4)$$

Where α is the learning rate. The parameters of NMT are guaranteed to converge into a local optima with a well-adjusted learning rate. In reality, adaptive learning rate optimizers like Adam [21] are found to significantly reduce training time compared to basic SGD optimizer.

Data augmentation

Overview

Data augmentation refers to techniques for increasing training data diversity without collecting extra data. Most methods either produce synthetic data or add slightly modified versions of existing data, expecting that the augmented data can serve as a regularizer and lessen overfitting while training machine learning models [22,23]. DA has been widely employed in computer vision, where model training typically includes operations like cropping, flipping, and color transforming. In NLP, where the input space is discrete, it is less evident how to create efficient augmented instances that capture the desired invariances.

Goals and trade-offs

Many DA strategies for NLP, from rule-based manipulations [11] to more complex generative systems [24], have been developed despite challenges associated with the text. Since the goal of DA is to offer an alternative for collecting more data, the optimal DA technique should be simple to use while also enhancing model performance. Most offer trade-offs between these two.

Rule-based methods are simple to apply but typically result in small enhancements in performance [7,25]. Conversely, approaches that utilize trained models may require more resources to implement but introduce greater data variability, resulting in more substantial performance improvements. Tailored model-based techniques for specific tasks can significantly impact performance but can be challenging to develop and utilize effectively.

Additionally, it is important that the augmented data distribution strike a balance between being neither too similar nor too different from the original data. If the augmented data is too similar, it may result in overfitting. At the same time, if it is too different, it can lead to poor performance due to training on examples that do not accurately represent the intended domain. Therefore, effective data augmentation approaches should strive for a harmonious equilibrium.

Discussing the interpretation of DA, authors from [26] note that ”data augmentation is typically performed in an ad-hoc manner with little understanding of the underlying theoretical principles” and claim that it is insufficient to explain DA as regularization. In general, there is a noticeable absence of comprehensive research on the precise mechanisms underlying the effectiveness of DA. Existing studies mostly focus on superficial aspects and seldom delve into the theoretical foundations and principals involved.

Techniques

The general approaches of DA techniques are mentioned in the survey of [27], the DA techniques can be grouped as rule-based techniques, interpolation techniques, and model-based techniques. For rule-based techniques, these techniques do not require model components and use simple, preset transforms. A typical example of these methods is EDA [7], which performs a set of random perturbation operations on the token level such as random insertion, swap, and deletion. A different category of DA techniques, called interpolation, initially introduced by MIXUP [28], involves interpolating the inputs and labels from multiple real examples. With the model-based approach, these DA techniques from this approach have also utilized Seq2seq models and language models. An example is the widely used ”back-translation” approach [9], which involves translating a sequence into a different language and then translating it back to the original language.

However, DA approaches can be more specifically categorized based on the characteristics of the techniques and the diversity of augmented data. The survey from [29] frames DA methods into three categories, including paraphrasing, noising, and sampling.

- The paraphrasing-based methods generate augmented data that retains a strong semantic resemblance to the original data by making controlled and limited modifications to the sentences. The augmented data effectively conveys almost identical information as the original data.
- The nosing-based methods aim to enhance the model’s robustness by introducing discrete or continuous noise while ensuring the validity of the data. These methods focus on adding noise in a controlled manner to improve the model’s ability to handle different scenarios.
- The sampling-based methods excel at understanding the data distributions and generating novel data samples from within these distributions. By employing artificial heuristics and trained models, these techniques produce more diverse data that effectively caters to a wider range of requirements for downstream tasks.

Method stacking

The augmentation methods above are not restricted to being used independently. They can be combined to achieve improved performance. Some common combinations include:

- The Same Type of Methods: Certain studies integrate various approaches based on paraphrasing and generate diverse paraphrases to enhance the diversity of augmented data. For example, the method in [30] uses both thesaurus and semantic embedding. Regarding methods based on noising, they often combine different techniques that were previously considered unlearnable, as demonstrated in research of [31]. This is because these methods are straightforward, efficient, and mutually beneficial. Some methods also adopt different sources of noising or paraphrasing like approach in [32]. The combination of different resources could also improve the robustness of the model.
- Unsupervised Methods: In certain situations, there is a need for straightforward and task-agnostic unsupervised data augmentation methods. Consequently, these methods are grouped together and extensively utilized. EDA is a very popular method that consists of synonym replacement, random insertion, random swap, and random deletion.
- Multi-granularity: Certain studies employ the same approach at various levels to enhance the augmented data by introducing diverse changes of varying granularities. This practice aims to enhance the model’s robustness. For example, the authors from [6] train both word embeddings and frame embeddings by Word2Vec.

BAHNAR LANGUAGE

Overview

Bahnar language or Ba-na language is a Mon-Khmer language belonging to the Bahnaric group and used by Bahnar people mainly living in central Vietnam [33]. Bahnaric can be divided into two sub-group: Northern Bahnar(Xêđăng, Halăng, Jeh,...) and Southern Bahnaric(Koho, Mnông, Chrau Jro,...). Bahnar language is an intermediate language of these two groups. Bahnar language is similar to Southern Bahnaric. Besides, the structure of its phonemes is simpler than the Northern Bahnaric. However, it shares more standard features in the vocabulary with Northern Bahnaric. Therefore, the Bahnar language can be classified as Central Bahnaric.

Vietnamese-Bahnar has unique differences from other translation pairs. From the perspective of Vietnamese-English translation, grammar plays a major role that requires the conversion of words into different variants depending on tense, active or passive voice, and in some specific cases, words or phrases may express different meanings which is the cultural characteristic of the language, slang for example. In terms of translation into languages containing pictograms (Chinese), information-rich and diverse word meanings require a high ratio of purposeful compression while preserving the correctness of the mentioned entities. In Vietnamese-Bahnar translation, in addition to the above problems, although there are similarities in sentence structure between the two languages, the existence of many sub-syllables with various accents, rare subsyllables, and cultural slang result in challenges for learners and also machine translation algorithms.

To be more specific, in terms of complexion, the word structure of the Bahnar language has several solid rules that connect firmly with lexemes. Words in the Bahnar language can be constructed using "affix", "reduplication", and "compound" [34]. Method affix is the most complicated one among the three ways. These methods can create derivational words with typical features such as variation word meanings and changing grammar functionality (e.g., nouns can become verbs).

Grammar rules

Bahnaric Vocabulary

Firstly, according to the Bahnaric language training program [35], the Bahnar language has a different character set than Vietnamese, containing 42 characters in total, with various accents. The above 42 characters form multiple-word components:

42 characters: a, ă, â, *b*, b, *čd*, d, e, ě, ê, ê, g, h, i, $\overset{\vee}{i}$, j, k, l, m, $\widetilde{n}$, o, $\overset{\vee}{o}$, ô, ô, o, $\overset{\vee}{o}$, p, r, s, t, u, $\overset{\vee}{u}$, u, $\overset{\vee}{u}$, w, y, f, q, v, x, z.

12 common diphthongs: ia, iă, ie, iě, iô, iô, uă, uă, ue, uě, uê, uê.

52 common consonants: bl, br, by, čh, čr, dj, djr, dr, gl, gr, gry, gy, hl, hm, hml, hmr, hmy, hn, hng, h$\widetilde{n}$, hr, hy, jr, kh, khy, kl, kr, ky, ly, ml, mr, ny, my, nr, ng, ngl, ngr, 'ngr, nhr, ngy, ny, ph, phr, phy, pl, pr, py, sr, th, thy, tr, ty.

Some of these consonants are scarcely used in real-world sentences, which is a challenge for Bahnaric learners: by, ly, ky…

Another considered syntactic point is the use of "'" before consonants and pre-syllable (sub-syllable or weak syllable, which can be considered sesquisyllabic) to construct different word forms but unchanged in meanings. Some typical pre-syllables can be listed as a, bơ, dơ, hơ, jơ, etc (Example: ame (chăm nom), bơbah (cuối nguồn), hơhoi (không có), jơnap (đầy đủ), etc).

From the view point of word form, similar to Vietnamese, Bahnaric words have singleword and compound forms, consisting of complex forms (words containing two or more syllables, each syllable is a meaningful single word), alliterative expressions such as adal adal (khe khẽ), kueng kueng (ẩm ẩm) or ring ring (nhộn nhịp).

Additionally, Bahnaric words can express different meanings in daily communication situations. For example, the word "adal adal" may mean soft when talking about voice or quietly making some actions, this word also means slowly and carefully or sometimes unhurriedly. As mentioned in 3.1, region and culture can affect vocabulary, which forms regional synonyms, resulting in suffering when mapping Bahnaric from one region to another.

Last but not least, Bahnaric vocabulary also contains loan words (mostly from Vietnamese), constructed by removing accents from Vietnamese words. Words in this vocabulary set are practically entities, especially named entities of persons, locations, organizations, and a small set of nouns and verbs.

Sentence Structure

The sentence structure of Bhanar is similar to that of Vietnamese. An ordinary simple sentence consists of two main components Subject and Predicate. The order of subject and predicate is the same as in Vietnamese.

Subject // Predicate

More complex sentences can be formed by combining various simple subject-predicate structures with conjunction (nhưng - mă lĕi - but, và - sơm - and, vì... nên... - yuô... kơna... - because... then..., ...). In real-life usage, the subject part in sub-sentences may be reduced or replaced by pronouns, which leads to a co-reference resolution problem to fully understand the whole sentence and make the correct translation.

Vietnamese-Bahnar translating notices

Translating from one language to another requires careful consideration of numerous factors that significantly influence the accuracy and quality of the translation.
These factors hold true when undertaking the translation process from Vietnamese to Bahnar as well, demanding ken attention and careful handling.

- Spelling: A spelling error refers to a deviation from the standard or accepted way of spelling a word. While it is not a major issue, it could occur due to errors in the input files.
- Collocation: Concerning the question of whether a specific phrase consists of words that naturally occur together or co-occur.
- Grammatical: Grammatical error refers to an occurrence of incorrect, unconventional, or disputed language usage, such as a misplaced modifier or an inappropriate verb tense.
- Typo: A typographical error, commonly known as a typo, refers to a mistake made during the typing or printing of written or electronic material. The most common typographical error comes from Bahnaric's unique and diverse accents along with the combination of presyllables and "" characters.
- Word-by-word translation: Word-for-word translation is commonly understood as the process of translating text from one language to another by directly using the exact words from the original text. This issue can create grammatical issues and collocation issues because of the complexity of regional word form and daily communication circumstances.

Besides, in normal conversation, there exist some different points between Bhanar and Vietnamese.

- Some sentences in Bahnar tend to skip words in simple sentences. For example, "Ở Vịnh Thạnh là đông nhất" in Vietnamese can be translated to "Uei Vinh Thanh lư loi"; in this case, the word "là" in Vietnamese can be skipped during translating.
- The position of exclamation in Bhanar is also different from Vietnamese. In Vietnamese, exclamation words usually stay behind the predicate, but it is the opposite in Bahnar. For example, "Mẹ ơi" in Vietnamese will be translated to "Ơ Mi" in Bahnar.

RELATED WORKS

Data augmentation in NMT

Overview

In recent years, the utilization of DA in these tasks has witnessed a notable increase [29]. Text classification, being the pioneering task to adopt DA, has garnered a larger number of research papers compared to the other two tasks: text generation and structure prediction. In Section "Data augmentation", it has been noted that each specific DA method has the potential to be implemented in text classification tasks. DA methods, which apply to text classification, can also apply to neural machine translation. However, due to the different nature of these tasks, some methods, which have shown powerful improvements in text classification, cannot perform well in neural machine translation. EDA is an example of the above statement. This method has created the abnormal in the context of sentences, such as: producing new vocabulary, changing word order, and skipping words. EDA will be presented in the next section and analyzed further in Section "Results and discussion" to prove its effectiveness. So, DA in NMT needs suitable approaches that are still based on the foundation of the original DA methods but need novelty modifications. Section "Related works" will state the recent DA methods applied to solve this problem. The methods will be categorized according to their characteristics. Figure 2 categorizes the typical methods in this section according to their respective DA approaches.

Related works

Numerous strategies for NLP, ranging from rule-based manipulations [36] to sophisticated generative systems [24], have been devised despite the challenges associated with working with text.
The back-translation approach is widely recognized as a popular DA method for NMT [37]. However, the approaches discussed in this section primarily concentrate on methods that do not rely on additional resources apart from the available training parallel corpus.
The study by [38] conducted an evaluation of back-translation and forward translation in a similar context. They trained NMT systems in both forward and

backward directions using the existing parallel data and then utilized these models to generate synthetic samples by translating either the target side (following the approach of[37]) or the source side (following the approach of[39]) of the original training corpus.

The two approaches are Reward Augmented Maximum Likelihood (RAML)[40] and its extension to the source language called SwitchOut[10]. These methods aim to expand the support of the empirical data distribution while maintaining its smoothness, ensuring that similar sentence pairs have similar probabilities. They achieve this by replacing words with other words sampled from a uniform distribution over the vocabulary. This approach tends to overstate infrequent words in practice. Additionally, the researchers from[41] proposed a related approach that promotes constituent behavior, where replaced words are selected from another sentence rather than from the vocabulary.

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, the research of[42,43] applied the technique of replacing tokens with placeholders, specifically in the source language. They combined this with auxiliary tasks that involved detecting replaced and dropped tokens. Similarly, the authors from[32] evaluated the impact of replacements on the target data but did not follow an MTL approach. Another related approach is word dropout, which has been explored by[44] and[45].

In terms of altering word order, there have been several proposals. The research by[46] and[47] have put forward their respective strategies. However, it is notable to mention the approach suggested by[48], 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. Random word replacement can be employed on the source side of training samples[32]. Replacing randomly selected words with soft words, whose representations are derived from the probability distribution provided by a language model[49]. A study by[7] has also applied this approach in EDA where they randomly replaced n words with their synonym. The authors of[50] replace several words in their training samples with infrequent words to enhance the NMT model's performance when translating such words. They identify words to be replaced using a large source language model and then use a word-alignment model and a probabilistic dictionary to replace the corresponding counterpart with the most probable translation of the new source word.

In the context of back-translation, the scientists from[51] experimented with various straightforward transformations such as word deletion, replacement, and swapping on the backtranslated data, resulting in a noticeable improvement. Regarding the special token used to prevent negative transfer between tasks, the researchers from[52] propose a similar approach to identify synthetic samples when combining actual parallel data and back-translated data for training. Approach of[53] expand upon this idea by incorporating forward-translated data for training and utilizing two distinct special tokens to distinguish between the two types of synthetic data.

Discussion

According to empirical research, back-translation and word replacement 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[54]. 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 the 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, none of them have ever applied in NMT DA to produce visible results from augmented data like text classification. Inspired by EDA and the multi-task learning approach[14], a multi-task learning framework with multiple operations should be conducted and examined to assess its effectiveness with individual methods. Besides, a further approach to the nosing-based method at the sentence level should be investigated. Therefore, combining truncated sentences on both the source and target sites beyond the word level could provide a promising way of applying the noising-based method to low-resource NMT.

APPROACHES

To develop dependable NMT systems, a substantial volume of parallel sentences is required. These parallel sentences consist of pairs of sentences in two

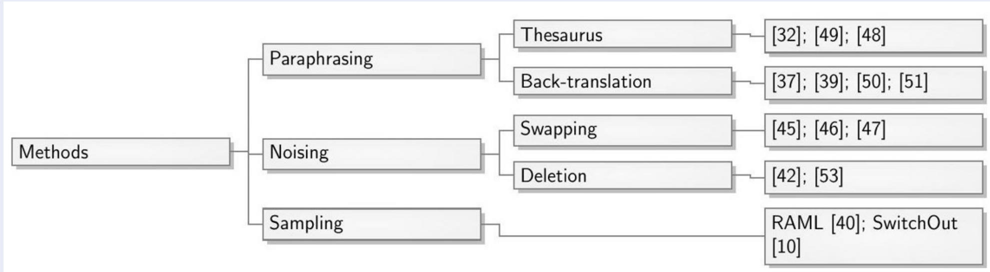

Figure 2: Taxonomy of DA methods

languages that are translations of each other. However, this poses a significant challenge for language pairs with limited resources. Building an NMT system for translating Vietnamese-Bahnar is a representative example of this challenge due to the extremely low-resource nature of Bahnar materials. Therefore, data augmentation techniques are involved in generating additional training samples when the available parallel data are scarce. To utilize the limited resources, the researcher applied some simple rule-based approaches to improve translation quality by augmenting the data. Two methods were proposed:

- Multi-task Learning Data Augmentation (MTL DA)
- Sentence Boundary Augmentation

Multi-task learning data augmentation (MTL DA)

This framework was inspired by EDA [7] and utilizes the DA operations that refer from [14]. The approach combines a set of simple DA methods to produce synthetic target sentences to strengthen the encoder. This framework is a set of word-level augmentation methods. These methods introduce the network to novel scenarios during training where relying solely on the target language context is insufficient to achieve a minimal loss. Consequently, the responsibility is shifted to the encoder. Recent findings by [55] further support this approach: they argue that the influence of source tokens in the output predictions of an NMT system decreases as decoding progresses. Moreover, this approach follows simple multi-task learning that does not require changes in the model architecture. MTL DA highlights that augmented data do not follow the distribution of a parallel corpus but can still produce a positive effect.

This framework does not require preprocessing steps, training additional systems, or data besides the available training parallel corpora, which makes it suitable for the low-resource target language, Bahnar. The framework includes five auxiliary tasks. For each task, a synthetic corpus of the same size as the original training data, obtained by transforming each original pair of sentences, will be appended to the training set. In almost all the tasks, the source sentence is left unchanged, while the target sentence is substantially modified.

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. Moving forward, t represents the count of words in the initial target sentence. Table 1 shows an example of the effect of different methods on a single sentence pair (hyperparameter α is set to 0.5 for three methods: Swap, token, and replace).

- Swap: Pairs of random target words are swapped until only $(1\text{-}\alpha)\cdot t$ remains in their original position. The objective of this task is to encourage the system to rely less on the target prefix when generating a new word.
- Token: Replacing $\alpha\cdot t$ random target words with a special (UNK) token [32]. With the same motivation as Swap, when generating a new word, the target prefix should become less informative and force the system to pay more attention to the encoder.
- Source: The target sentence is transformed into an identical copy of the source sentence. As a result, the optimal strategy for generating the correct output is to rely on the encoder representation and replicate the relevant information from the source.
- Reverse: The order of the words in the target sentence is reversed. According to [55], the impact of the encoder diminishes as the target sentence progresses. Hence, by reversing the sentence order, the system is expected to learn to rely more

on the encoder for generating words that typically occur towards the end of the sentence, utilizing additional information.

- Replace: $\alpha \cdot t$ source-target aligned words are selected at random and replaced by random words in the dictionary. Unlike[50] followed constrained replacements to produce only fluent target sentences, this modification is expected to introduce challenging words that cannot be easily generated solely based on the target language prefix. As a result, the system is compelled to focus on the source words and pay closer attention to them.

The operations above will be strategically combined based on the classic stacking method of data augmentation. This approach aims to achieve optimal results with care- fully selected augmented datasets. In contrast to the EDA method, which incorporates all four of its operations within a unified framework, our methodology does not aggregate all operations indiscriminately. Instead, our focus is on the thoughtful selection and integration of the most appropriate methods, guided by empirical experiments and evaluations. This principle has been substantiated in Section "Results and discussion".

Sentence boundary augmentation

In this approach, the very original key idea is applying sentence-level augmentation to low-resource NMT. This approach was originally motivated by the proposal of[15]. Their research has proved that sentence boundary augmentation has successfully overcome the issue of wrong segmentation for translating sentences produced by automatic speech recognition (ASR) systems. Moreover, study by[56,57] have also indicated that translation degradation is caused by poor system sentence boundary prediction. It is reasonable to assume that sentence boundary errors could impact translation quality. Incorrect boundaries might separate words from crucial contextual information necessary for accurate translation. When provided with sufficient context, the system could make mistakes simply due to unexpected placements of sentence boundaries. This method effectively addresses the problem of incorrect segmentation present in the limited resources of the dataset on the Bahnar language side. Besides, this method was supported by the research of[58]. This research has stated a significant point of augmenting data on the sentence level, although this research is in a different context. The researcher applied the noising-based method at the sentence level in the context of sentiment analysis for Spanish tweets. Although divided tweets were split into two halves and combined randomly sampled first and second halves with the same label, they 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.

Algorithm 1 Sentence Boundary Augmentation[15]

Input: List of source S, target T as tuple $L = \{(S_i, T_i)\}_{i=0}^{n}$, hyper-parameter $p = 0.3$ for sentence length truncation

1: $L_{out} \leftarrow$ list();
2: **for** i = 0,2,4,…,n-2,n **do**
3: $p \sim Uniform(0, p)$;
4: $(S_1, T_1),(S_2, T_2) = L[i], L[i+1]$;
5: $p_1^{s_1}, p_1^{t_1} = [p \times len(S_1)], [p \times len(T_1)]$;
6: $p_1^{s_2}, p_1^{t_2} = [p \times len(S_2)], [p \times len(T_2)]$;
7: $S_{out} = concat\left(S_1\left[p_1^{s_1}:\right], S_2\left[:p_2^{s_2}\right]\right)$;
8: $T_{out} = concat\left(T_1\left[p_1^{T_1}:\right], S_2\left[:p_2^{T_2}\right]\right)$;
9: $L_{out}.append([S_{out}, T_{out}])$;
10: **end for**
11: **return** L_{out}

The pseudocode for this approach is given in Algorithm 1, which represents sentences as sequences of tokens. First, the list of pair sentences is fed to the algorithm as input. In the initial step (line 1), the algorithm initializes an empty list L_{out} to store the augmented sentence pairs. In the looping process from line 2 to line 10, every two continuous sentence pairs are processed to produce a new augmented sentence pair. In line 3, portion p will follow a uniform distribution of 0 and a defined hyperparameter p (p=30% by default). (S_1, T_1),(S_2, T_2) are two continuous sentence pairs where S_1, T_1 are the first source and target sentence, and S_2, T_2 are the next second source and target sentence (line 4). Next, all four partitions are determined based on the ceiling value of the multiplication between the length of each respective sentence and portion p. The two values $p_1^{S_1}$, $p_1^{T_1}$ are the partitions of the first source sentence and the first target sentence, respectively (line 5). In line 6, $p_2^{S_2}$, $p_2^{T_2}$ are calculated similarly to previous partition values, but only with the second sentence pair. Adjacent source sentences S, and their corresponding target sentences T are concatenated. Next, the start and end of the concatenated source and target sentences are proportionally truncated to imitate a random start or break. This is governed by the computed partition values. By design, the truncation keeps most of the first sentence (lines 7,8 for S_1, T_1: at most 30% of the start of the first sentence is discarded) while discarding the bulk of the second sentence (lines 7,8 forS_2, T_2: at most 30% of

Table 1: Example of Multi-task Learning Data Augmentation.

Task	Language	Augmented sample
Original training sample	source	Bố Điêu bị ốm nặng
	target	Bă đe Diêu jǐ adrin
Swap	target	Diêu đe Bă jǐ adrin
Token	target	Bă đe UNK jǐ UNK
Source	target	Bố Điêu bị ốm nặng
Reverse	target	adrin jǐ Diêu đe Bă
Replace	source	con vẹt Điêu vàng ốm nặng
	target	sem diê ’brơu Diêu jǐ adrin

the start of the second sentence is retained). This removes context from the first sentence S_1, adds context from the second sentence S_2, and combines them into a single training example. That new augmented sample will be pushed into the L_{out} in line 9 . After the algorithm passes through all sentences, it will return the list of augmented sentence pairs in line 11. Table 2 shows how truncated sentences are formed to create an augmented sentence.

In the original research, the authors may have solely focused on the algorithm. However, we aim to delve deeper into the effectiveness, especially concerning any changes in the parameter p, to ensure that the new augmented data does not diverge significantly, despite potential variations in the value of p.

Pipeline

To investigate the effect of DA on translating performance, the training set will be augmented with a DA module that contains several DA methods, including MTL DA and sentence boundary augmentation. Besides, two more methods, EDA and semantic embedding, will also be experimented to analyze the effects. The new augmented training set is a combination of the original training data and the new augmented data. With each different DA approach, there will be a unique version of the augmented data set, and each of them will be passed through the same training process. In other words, this pipeline was designed to generate a number of new augmented datasets corresponding to each applied data augmentation method. After training, the results of the augmented data will be collected, compared, and evaluated to see the performance of each method's impact. Figure 3 shows the general approach of how the researcher conducts this project.

EXPERIMENTS AND DISCUSSION

Dataset

The dataset used for this project is the bilingual corpus of Vietnamese-Bahnar. The dataset contains a thousand lines of text in both Vietnamese and Bahnar. The dataset's sentences were initially collected from online news. The translated Bahnar version is made manually by the Bahnar collaborators based on the Bahnar Kriem dictionary. Later, the dataset was improved with more sentence pairs; these pairs were collected from the teaching textbook. These sentences are formal greetings, formal and informal conversations, narrative stories, and folktales written in Bahnar Kriem.

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. Each sub-dataset contains two text files; the first text file is used for storing Vietnamese sentences (stored as .vi), and Bahnar sentences are stored in the other text file (stored as .ba). The total number of sentence pairs in each sub-set is shown in Table 3.

Table 3: Original Dataset

Type	# of sentence pairs
Training set	16105
Test set	1988
Valid set	1987

Evaluation metric

To evaluate experimental results, the main metric that is used for evaluating is BLEU (Bilingual Evaluation Understudy)[59]. BLEU is a metric for automatically evaluating machine-translated text. The BLEU

Table 2: Example of Sentence Boundary Augmentation applying for both Vietnamese and Bahnar

Sentences	Source Sentence	Target Sentence
S_1, T_1	Phó Trưởng Ban thường trực: Ông Phan Trọng Hổ, Giám đốc Sở Nông nghiệp và Phát triển nông thôn	Phǒ Trương 'Ban thương trưk: 'Bok Phan Trong Hô, Giam đôk Sơ Nông nghiêp weng pơjing cham pơlĕi
S_2, T_2	Vì vậy, ngành y tế huyện, khuyến cáo người dân thận trọng trong việc sử dụng các loại nấm, tuyệt đối không được sử dụng các loại nấm lạ, để tránh bị ngộ độc	Yua noh, nganh y tê hŭn pơtho khan nă ma wă băt lơm tơdrong chă yuô rim loai mơu, pơgloh bi đĕi chă yuô rim loai mơu la sư hli ngô đôc
S_{out}, T_{out}	Ông Phan Trọng Hổ , Giám đốc Sở Nông nghiệp và Phát triển nông thôn. Vì vậy , ngành y tế huyện	'Bok Phan Trong Hô , Giam đôk Sơ Nông nghiêp weng pơjing cham pơlĕi Yua noh , nganh y tê hŭn

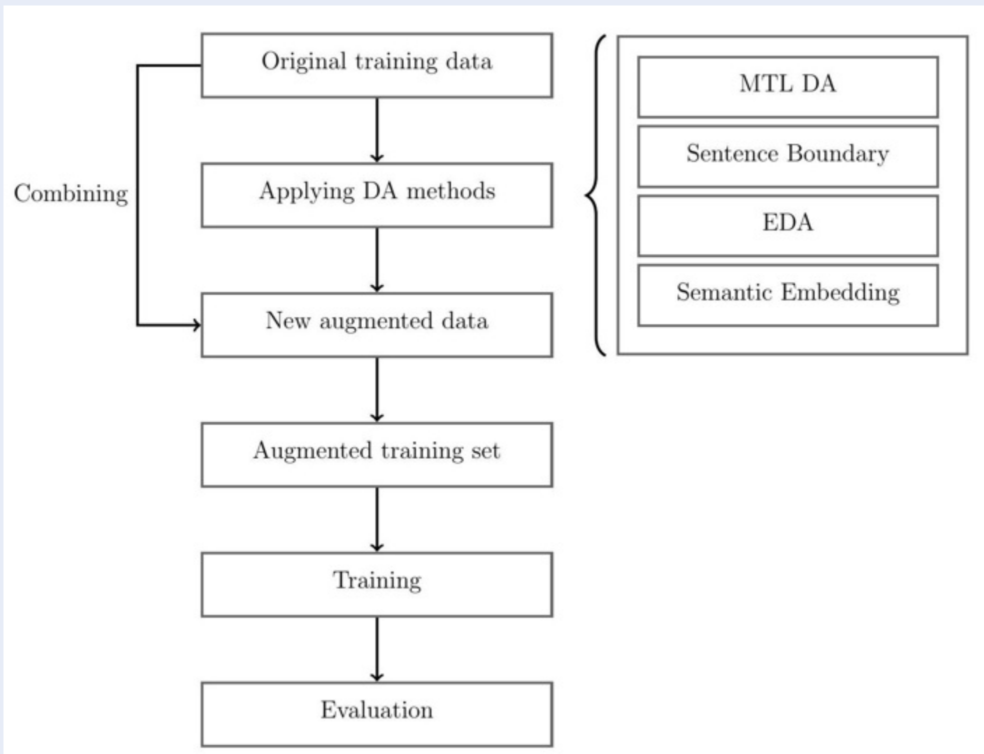

Figure 3: General pipeline of augmenting, training and evaluating process

score ranges from 0 to 1, representing how closely the machine-translated text resembles a set of excellent reference translations. Number 0 indicates little to no overlap between the machine-translated output and the reference translation (poor quality), whereas 1 indicates perfect overlap between the two translations (high quality). Sometimes, the decimal values are turned into a 0 to 100 scale so they can be read more easily; for example, 0.7 can be written as 70 .

It has been demonstrated that BLEU scores and human evaluations of translation quality are highly correlated. It should be noted that even human translators do not get a score of 1.0. It is strongly advised against comparing BLEU scores across different corpora and languages. Even comparing BLEU scores within the same corpus but with varying numbers of reference translations can lead to highly misleading results. However, as a rough guideline, Table 4 shows the interpretation of BLEU scores.

Table 4: Interpretation of BLEU scores [60]

BLEU Score	Interpretation
< 0.10	Almost useless
0.10 - 0.19	Hard to get the gist
0.20 - 0.29	The gist is clear, but there are substantial grammatical errors present
0.30 - 0.40	Understandable to good translations
0.40 - 0.50	High quality translations
0.50 - 0.60	Very high quality, adequate, and fluent translations
> 0.60	Quality often better than human

Experimental settings

To experiment, Transformers[a] was chosen as the core framework. The framework compatibility of JAX, TensorFlow, and PyTorch is supported through transformers. With this flexibility, a model can be trained in one framework with a few lines of code and loaded for inference in another, depending on the stage of its existence. Google Colab[b] was chosen as an environment for training the model.

The one-to-one word alignments required by replace of MTL DA were obtained by using SimAlign[c]. This alignment mechanism leverages multilingual word embeddings - both static and contextualized for word alignment [61]. Multilingual embeddings are created from monolingual data only without relying on parallel data or dictionaries. One of the proposed default embedding models has been chosen for the embedding model used for the aligner, xlm-mlm-100-1280.

During the training process, the author only tried to train and test the translation of Vietnamese-Bahnar using the model BARTpho. DA methods were studied and experimented with to see their effectiveness on translation performance. In the training process, $BART_{phosyllable}$ was chosen as the pre-trained and baseline model, and the BLEU score is the primary metric to evaluate the best training model. All models were trained with the hyperparameters shown in Table 5.

Table 5: Training Hyperparameters

Parameter	Value
Training batch size	32
Evaluation batch size	32
Epoch	2
Learning rate	2e - 5
Weight decay	0.01
Evaluation steps	1000
Num beams	5
Valid set	1987

Results and discussion

General

To conduct the experiments, several augmented training sets were generated separately. With MTL DA and sentence boundary method, the size of generated data points was consistent. While MTL DA is twice as large as the original dataset size, sentence boundary DA is only 0.67 times that of the origin (Shown in Table 6). The augmented data size of the other two methods (EDA and semantic embedding) is followed by the exact value of MTL DA to evaluate fairly.

Table 6: Total sentence pairs of the baseline and augmented training sets

Method	# of sentence pairs
Baseline	16105
MTL-DA	32210
Sentence boundary	24157
EDA	32210
Semantic embedding	32210

Table 7 reports the translation performance, measured in terms of BLEU for prediction. In this experiment, hyperparameter α was chosen as 0.5 because if α is too small, nearly none of the transformations can apply to target sentences in the training set. So, augmented data will be very similar to original data, which either violates diversity criteria or harms future models in the out-domain fields. In contrast, if α is

[a] visit https://github.com/huggingface/transformers
[b] visit https://colab.research.google.com/
[c] https://github.com/cisnlp/simalign

too large, the translating model of the in-domain can suffer from poor performance.

Table 7: BLEU scores of baseline and MTL DA approach, using different auxiliary tasks and their combinations

Method	BLEU
Baseline	29.89
Swap	35.22
Token	**38.48**
Source	2.38
Reverse	31.14
Replace	**38.01**
Replace+swap	37.91
Replace+token	38.35
Token+swap	**40.64**
Replace+token+swap	37.03

First of all, the baseline is the evaluation results when training and testing without any augmentation method applied. The results show that the MTL DA approach consistently outperforms the baseline system except for method source. Generally, the auxiliary tasks token and replace are the best-performing ones. swap and reverse may give a lower performance result than these two methods above, which suggests that abnormal word order could negatively influence the main task, but it still produces a promising result. While all five methods have shown improved results, source has its worst performance, which indicates that the translation task could be adversely affected by introducing a completely different vocabulary in the target.

Interestingly, using each two of the three best auxiliary tasks together further improves the performance, achieving well-performed results in all translation tasks with BLEU scores between 8.02(replace+swap) and 10.75 (token+swap) points over the baseline. The combination of token+swap gives the best performance. Although the researcher has combined all three best methods, the results still cannot outperform the combination of token and swap. In general, all combinations have led to an enhancement in the BLEU score. This suggests that different auxiliary tasks have unique effects on the encoder and show a form of mutual support.

In sentence boundary augmentation, DA performs based on truncated sentence combination; this combination depends on hyperparameter p. Therefore, the researcher has examined different values of p on the results (shown in Table 8). Surprisingly, the values of p do not affect so much on translation results. The BLEU scores of different p values are slightly different, although, in this specific project, p=0.7 has shown the most well-performed results compared with others. This method has shown a significant advance compared to the baseline system.

Table 8: BLEU scores obtained using different p in sentence boundary augmentation approach

p	BLEU
$p = 0.1$	40.34
$p = 0.3$	39.86
$p = 0.5$	40.47
$p = 0.7$	**41.33**
$p = 0.9$	40.86

Considering the same training configuration, with the less generated sentence pairs, the best result of sentence boundary augmentation can be on pair 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 at least. It can also utilize limited resources to produce a smaller augmented training set. For low-resource machine translation augmentation, the nosing-based method applied at the phrase or sentence level has demonstrated superior performance to the nosing-based method applied at the word level.

The researcher chooses hyperparameter α=0.5 for EDA and semantic embedding. Both methods were applied on the target site to serve the same purpose of MTL DA: to strengthen the encoder and make a word dependent less on the prefix word. Four methods can overcome the baseline, but EDA has fallen behind the other three methods (shown in Table 9). This issue happens due to such reasons: random insertion and random replacement need to use the external resource from the dictionary, which may provide some new words to the vocabulary; random swap and random deletion could accidentally change the word; these four methods combined together could create a negative effect. Therefore, EDA, initially designed for text classification, can not be utilized in the context of NMT. Semantic embedding may have good performance but still cannot compare with MTL DA combination and sentence boundary, which can prove that word replacement is a possible and promising solution, but it needs to have reasonable strategies to decide how the method should be used, such as augmented language site, hyperparameter, utilizing language model, strategy in choosing words to replace, etc.

Table 9: BLEU scores obtained with baseline, MTL DA, sentence boundary, EDA, and semantic embedding

Method	BLEU
Baseline	29.89
EDA	36.37
Semantic embedding	39.20
Token+swap	40.64
Sentence boundary	41.33

Sentence orientation

Section "Vietnamese-Bahnar translating notices" has mentioned several issues that might happen during the translating process, such as collocation issues, word-by-word translating issues, etc. Therefore, this project will also focus on improving these specific cases. However, it is impossible to cover all the mentioned issues in the whole testing set; only a set of sentences is chosen to investigate. Based on suggestions from Table 4, the researcher chose the translated sentences with a BLEU score range from 0.2 to 0.4 . Because it is complicated to detect translating issues with a low BLEU score sentence; on the other hand, a sentence with a high BLEU score is good, making it hard to exploit issues. Based on the given criteria, there are only 222 satisfied sentences from the test set. Table 10 shows each issue and its total sentences.

Table 10: Translating issues of chosen sentences in test set

Issue	# sentences
Collocation	34
Word-by-Word	36
Number ambiguity	102
Unknown	50

Collocation and word-by-word have been mentioned before, and the leftover is ”number ambiguity” and ”unknown”. ”Number ambiguity” refers to the issue that numbers are written as text in the test set, but their prediction is in number. Because using the BLEU score, which relies heavily on tokens, the BLEU score of these sentences is really low, although other parts are translated well. ”Unknown” can be understood as unstable in translating a long sentence that is hard to identify which issue occurs within. Therefore, in this section, experiments will only focus on two issues: Collocation and word-by-word.

Table 11 represents all the experiments of baseline and DA augmented models to two issues, ”collocation” and ”word-by-word”. In this experiment, the researcher only focuses on MTL DA(token+swap) and sentence boundary augmentation because both have proved efficient compared to baseline and other methods. In general, the predicted BLEU scores of ”collocation” are always smaller than ”word-by-word” in both baseline and augmentation contexts, which indicates that the collocation issue is more complicated to handle than the other one. Overall, two augmentation methods have significantly improved the above linguistic issues.

For collocation, the MTL DA-based method - ”token+swap” grows over three times higher than the baseline, while ”sentence boundary” is 2.5 times higher when compared with the baseline. On the other hand, for ”word-by-word”, ”sentence boundary” reaches almost 68% better performance when compared with the baseline; however, the ”token+swap” method can only achieve nearly 40% higher than the baseline. Interestingly, each context has its own dominant DA method. For collocation, the effective DA method was ”token+swap”, while ”sentence boundary” was the method that gave the best support to ”word-by-word”. There is a reasonable explanation for both cases. With word-by-word, the issue comes from the fraction translations, which keep the same original words; however, ”sentence boundary” can augment the data and produce more stable sentence-level truncation simultaneously. Therefore, the augmented models can learn from the new set of words. Noticing that ”sentence boundary” in ”collocation” has a lower performance than MTL DA, this issue happens because the focused collocations in the experiment are usually allocated at the beginning or the end of the sentence, which might be truncated due to the default algorithm of sentence boundary (see in Algorithm 1).

CONCLUSION

The main point of this paper is data augmentation for neural machine translation, especially data augmentation methods for translating low-resource languages, which is translating Vietnamese to Bahnar in this context. Several types of research have been studied; the DA methods have been addressed from word replacing, nosing-based method, to back-translation. Besides, the attributes of the Bahnar language have been studied to explore more linguistic features for

Table 11: BLEU scores of Collocation and word-byword with baseline and other DA methods

Method	Collocation	word-by-word
baseline	1.52	6.61
token+swap	4.64	9.19
sentence boundary	3.87	11.08

conducting this project. In this project, two approaches have been suggested and effectively implemented for both scenarios: general sentences and oriented sentences.

Firstly, Multi-task Learning Data Augmentation, known as MTL DA, differs from conventional approaches that focus on expanding the support of empirical data distribution by generating new samples that align with this distribution. These techniques produce new sentence pairs by applying radical transformations synthetically. The translation of augmented sentences introduces fresh contexts during training. When translating augmented sentences during training, the system learns to rely more on the encoder because the target prefix alone is insufficient to predict the next word accurately. As a result, the encoder is reinforced, and the system is compelled to rely more on it. Experiments were carried out on five low-resource translation tasks, four of which have improved over the baseline system. The other one (source) represents a harmful affection when adding a whole new vocabulary on the target side.

Secondly, Sentence Boundary Augmentation, which explores the noising-base method (swap) beyond word level by applying sentence level for augmenting. This is also a significant point when a method with one purpose (overcoming wrong segmentation for translation in the ASR system) can apply for another purpose (low-resource NMT), and it even makes an impressive performance result (improving the average 10 BLEU score compared with the baseline). This interference presents an opportunity for exploring and applying methods from other aspects to enhance low-resource NMT.

Both of these approaches are independent of the specific architecture of the NMT model. Furthermore, they do not necessitate complex preprocessing steps, training of additional systems, or acquiring additional data beyond the existing training parallel corpora. However, most of the proposed operations are implemented in the target language; only two methods are applied on both sides. Furthermore, it could be combined with existing DA methods, such as back-translation, especially those that operate on the source side [10,49].

Besides, further improvements could be achieved by implementing more sophisticated approaches to multi-task learning, such as changing the proportion of data for the different tasks and evaluating different ways of parameter sharing between the different tasks. Moreover, to evaluate the validity of these methods, different low-resource datasets (e.g. different dialects, different low-resource languages) should be trained and tested.

COMPETING INTERESTS

The author(s) declare that they have no competing interests.

AUTHOR CONTRIBUTION STATEMENT

- Sang T. Nguyen: Researching data augmentation methods in neural machine translation, proposing approaches, augmenting dataset, evaluating results, and writing research paper.
- Nguyen Q. Pham: Researching neural machine translation, Bahnaric language characteristics, supporting collecting resources, and writing research paper.
- Tho Quan: Come up with ideas for writing articles, providing dataset in Vietnamese, Bahnaric. Providing paper tutorials and editing.

ACKNOWLEDGMENT(S)

This research is funded by the 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-4.0-29/1925.

REFERENCES

1. Kalchbrenner N, Blunsom P. Recurrent Continuous Translation Models. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL. ACL; 2013. p. 1700-9;Available from: https://aclanthology.org/D13-1176/.

2. Sutskever I, Vinyals O, Le QV. Sequence to Sequence Learning with Neural Networks. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. NIPS'14. Cambridge, MA, USA: MIT Press; 2014. p. 3104-3112;.
3. Linh T. Biên soạn sách Tiếng dân tộc thiểu số chậm so với tiến độ, Bộ Giáo dục nói gì? Giáo Dục Việt Nam. 2021 August;Available from: https://giaoduc.net.vn/bien-soan-sach-tieng-dan-toc-thieuso-cham-so-voi-tien-do-bo-giao-duc-noi-gi-post220364.gd.
4. Kocmi T, Bojar O. Trivial Transfer Learning for Low-Resource Neural Machine Translation. In: Proceedings of the Third Conference on Machine Translation: Research Papers. Brussels, Belgium: Association for Computational Linguistics; 2018. p. 244-52;Available from: https://aclanthology.org/w18-6325.
5. Le Cun Y, Jackel LD, Bottou L, Cortes C, Denker JS, Drucker H, et al. Learning Algorithms For Classification: A Comparison On Handwritten Digit Recognition. In: Oh JH, Kwon C, Cho S, editors. Neural Networks: The Statistical Mechanics Perspective. World Scientific; 1995. p. 261-76;Available from: http://leon.bottou.org/papers/lecun-95a.
6. Wang WY, Yang D. 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. Lisbon, Portugal: Association for Computational Linguistics; 2015. p. 2557-63;Available from: https://aclanthology.org/D15-1306.
7. Wei J, Zou K. 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). Hong Kong, China: Association for Computational Linguistics; 2019. p. 6382-8;Available from: https://aclanthology.org/D19-1670.
8. Zhang D, Li T, Zhang H, Yin B. On Data Augmentation for Extreme Multi-label Classification. CoRR. 2020;abs/2009.10778;Available from: https://arxiv.org/abs/2009.10778.
9. Sennrich R, Haddow B, Birch A. 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). Berlin, Germany: Association for Computational Linguistics; 2016. p. 86-96;Available from: https://doi.org/10.18653/v1/P16-1009.
10. Wang X, Pham H, Dai Z, Neubig G. SwitchOut: an Efficient Data Augmentation Algorithm for Neural Machine Translation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics; 2018. p. 856-61;Available from: https://doi.org/10.18653/v1/D18-1100.
11. Zhang X, Zhao J, LeCun Y. Character-level Convolutional Networks for Text Classification. In: Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R, editors. Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc.; 2015;Available from: https://proceedings.neurips.cc/paper_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.
12. Chen Y, Lu S, Yang F, Huang X, Fan X, Guo C. Pattern-aware Data Augmentation for Query Rewriting in Voice Assistant Systems. CoRR. 2020;abs/2012.11468;Available from: https://doi.org/10.5465/AMBPP.2020.11468abstract.
13. Longpre S, Wang Y, DuBois C. How Effective is Task-Agnostic Data Augmentation for Pretrained Transformers? In: Cohn T, He Y, Liu Y, editors. Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020. vol. EMNLP 2020 of Findings of ACL. Association for Computational Linguistics; 2020. p. 4401-11;Available from: https://doi.org/10.18653/v1/2020.findings-emnlp.394.
14. Sánchez-Cartagena VM, Esplà-Gomis M, Pérez-Ortiz JA, Sánchez-Martínez F. 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. Association for Computational Linguistics; 2021. p. 8502-8516;.
15. Li D, I T, Arivazhagan N, Cherry C, Padfield D. Sentence Boundary Augmentation for Neural Machine Translation Robustness. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2021. p. 7553-7;.
16. Budiansky S. Lost in Translation. Atlantic Magazine. 1951 12:81-4;.
17. Jurafsky D, Martin JH. Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, N.J.: Pearson Prentice Hall; 2009;Available from: http://www.amazon.com/Speech-Language-Processing2nd-Edition/dp/0131873210/ref=pd_bxgy_b_img_y.
18. Brown PF, Cocke J, Pietra SD, Pietra VJD, Jelinek F, Lafferty JD, et al. A Statistical Approach to Machine Translation. Comput Linguistics. 1990;16:79-85;.
19. Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. In: Bengio Y, LeCun Y, editors. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings; 2015;Available from: http://arxiv.org/abs/1409.0473.
20. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All You Need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 6000-6010;.
21. Kingma D, Ba J. Adam: A Method for Stochastic Optimization. International Conference on Learning Representations. 201412;.
22. Hernández-García A, König P. Data augmentation instead of explicit regularization. CoRR. 2018;abs/1806.03852;Available from: http://arxiv.org/abs/1806.03852.
23. Shorten C, Khoshgoftaar T. A survey on Image Data Augmentation for Deep Learning. Journal of Big Data. 2019 07;6;Available from: https://doi.org/10.1186/s40537-019-0197-0.
24. Liu R, Xu G, Jia C, Ma W, Wang L, Vosoughi S. 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). Online: Association for Computational Linguistics; 2020. p. 9031-41;Available from: https://aclanthology.org/2020.emnlp-main.726.
25. Xu N, Mao W, Wei P, Zeng DD. MDA: Multimodal Data Augmentation Framework for Boosting Performance on Sentiment/Emotion Classification Tasks. IEEE Intelligent Systems. 2021;36:3-12;Available from: https://doi.org/10.1109/MIS.2020.3026715.
26. Dao T, Gu A, Ratner AJ, Smith V, Sa CD, Ré C. A Kernel Theory of Modern Data Augmentation. International Conference on Machine Learning. 2019 May;.
27. Feng SY, Gangal V, Wei J, Chandar S, Vosoughi S, Mitamura T, et al. A Survey of Data Augmentation Approaches for NLP. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online: Association for Computational Linguistics; 2021. p. 968-88;Available from: https://aclanthology.org/2021.findings-acl.84.
28. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. mixup: Beyond Empirical Risk Minimization. In: International Conference on Learning Representations; 2018;Available from: https://openreview.net/forum?id=r1Ddp1-Rb.
29. Li B, Hou Y, Che W. Data augmentation approaches in natural language processing: A survey. AI Open. 2022;3:71-90;Available from: https://www.sciencedirect.com/science/article/pii/S2666651022000080.
30. Liu S, Lee K, Lee I. Document-level multi-topic sentiment classification of Email data with BiLSTM and data augmen-

tation. Knowledge-Based Systems. 2020;197:105918;Available from: https://www.sciencedirect.com/science/article/pii/S0950705120302574.

31. Peng B, Zhu C, Zeng M, Gao J. Data Augmentation for Spoken Language Understanding via Pretrained Language Models. In: Proc. Interspeech 2021; 2021. p. 1219-23;Available from: https://doi.org/10.21437/Interspeech.2021-117.
32. Xie Z, Wang SI, Li J, Lévy D, Nie A, Jurafsky D, et al. Data Noising as Smoothing in Neural Network Language Models. In: International Conference on Learning Representations; 2017;Available from: https://openreview.net/forum?id=H1VyHY9gg.
33. Jil Y, H'MER, Đinh Văn H, Đinh Văn K. Quy tắc chính tả tiếng Bahnar. Sổ tay phương ngữ tiếng Bahnar. 2018;.
34. Nghia TD. Vấn đề hình vị trong tiếng Bahnar và các ngôn ngữ dân tộc thiểu số vùng Nam Bộ. Tạp Chí Khoa Học Xã Hội. 2009;9:50-4;.
35. Chương trình giáo dục phổ thông tiếng Bahnar. Ministry of Education and Training. 2020 September;Available from: https://moet.gov.vn/content/vanban/Lists/VBPQ/Attachments/1372/CT%20Tieng%20Bahnar.pdf.
36. Zhang X, Zhao JJ, LeCun Y. Character-level Convolutional Networks for Text Clas- sification. In: NIPS; 2015;Available from: https://api.semanticscholar.org/CorpusID:368182.
37. Sennrich R, Haddow B, Birch A. Edinburgh Neural Machine Translation Systems for WMT 16. In: Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. Berlin, Germany: Association for Computational Linguistics; 2016. p. 371-6;Available from: https://doi.org/10.18653/v1/W16-2323.
38. Li G, Liu L, Huang G, Zhu C, Zhao T. Understanding Data Augmentation in Neural Machine Translation: Two Perspectives towards Generalization. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP). Hong Kong, China: Association for Computational Linguistics; 2019. p. 5689-95;PMID: 30950606. Available from: https://doi.org/10.18653/v1/D19-1570.
39. Zhang J, Zong C. Exploiting Source-side Monolingual Data in Neural Machine Translation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: Association for Computational Linguistics; 2016. p. 1535-45;Available from: https://doi.org/10.18653/v1/D16-1160.
40. Norouzi M, Bengio S, Chen z, Jaitly N, Schuster M, Wu Y, et al. Reward Augmented Maximum Likelihood for Neural Structured Prediction. In: Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R, editors. Advances in Neural Information Processing Systems. vol. 29. Curran Associates, Inc.; 2016;Available from: https://proceedings.neurips.cc/paper_files/paper/2016/file/2f885d0fbe2e131bfc9d98363e55d1d4-Paper.pdf.
41. Guo D, Kim Y, Rush A. Sequence-Level Mixed Sample Data Augmentation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics; 2020. p. 5547-52;PMID: 32609793. Available from: https://doi.org/10.18653/v1/2020.emnlp-main.447.
42. Zhang H, Qiu S, Duan X, Zhang M. Token Drop mechanism for Neural Machine Translation. In: Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics; 2020. p. 4298-303;Available from: https://doi.org/10.18653/v1/2020.coling-main.379.
43. Shen D, Zheng M, Shen Y, Qu Y, Chen W. A Simple but Tough-to-Beat Data Augmentation Approach for Natural Language Understanding and Generation. CoRR. 2020;abs/2009.13818;Available from: https://arxiv.org/abs/2009.13818.
44. Gal Y, Ghahramani Z. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. NIPS'16. Red Hook, NY, USA: Curran Associates Inc.; 2016. p. 1027-1035;.
45. Arora S, Liang Y, Ma T. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In: International Conference on Learning Representations; 2017;Available from: https://api.semanticscholar.org/CorpusID:64908139.
46. Artetxe M, Labaka G, Agirre E. Unsupervised Statistical Machine Translation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics; 2018. p. 3632-42;Available from: https://doi.org/10.18653/v1/D18-1399.
47. Lample G, Ott M, Conneau A, Denoyer L, Ranzato M. Phrase-Based & Neural Unsupervised Machine Translation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics; 2018. p. 5039-49;Available from: https://doi.org/10.18653/v1/D18-1549.
48. Zhang Z, Wu S, Liu S, Li M, Zhou M, Chen E. Regularizing Neural Machine Translation by Target-Bidirectional Agreement. In: AAAI Conference on Artificial Intelligence; 2019;Available from: https://doi.org/10.1609/aaai.v33i01.3301443.
49. Gao F, Zhu J, Wu L, Xia Y, Qin T, Cheng X, et al. Soft Contextual Data Augmentation for Neural Machine Translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 5539-44;Available from: https://doi.org/10.18653/v1/P19-1555.
50. Fadaee M, Bisazza A, Monz C. 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). Vancouver, Canada: Association for Computational Linguistics; 2017. p. 567-73;Available from: https://doi.org/10.18653/v1/P17-2090.
51. Edunov S, Ott M, Auli M, Grangier D. Understanding Back-Translation at Scale. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics; 2018. p. 489-500;Available from: https://doi.org/10.18653/v1/D18-1045.
52. Caswell I, Chelba C, Grangier D. Tagged Back-Translation. In: Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers). Florence, Italy: Association for Computational Linguistics; 2019. p. 53-63;Available from: https://doi.org/10.18653/v1/W19-5206.
53. Yang Z, Chen W, Wang F, Xu B. Effectively training neural machine translation models with monolingual data. Neurocomputing. 2019;333:240-7;Available from: https://doi.org/10.1016/j.neucom.2018.12.032.
54. Chapelle O, Scholkopf B, Zien A Eds. Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006) [Book reviews]. IEEE Transactions on Neural Networks. 2009;20(3):542-2;Available from: https://doi.org/10.1109/TNN.2009.2015974.
55. Voita E, Sennrich R, Titov I. Analyzing the Source and Target Contributions to Predictions in Neural Machine Translation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics; 2021. p. 1126-40;Available from: https://doi.org/10.18653/v1/2021.acl-long.91.
56. Gangi MAD, Enyedi R, Brusadin A, Federico M. Robust Neural Machine Translation for Clean and Noisy Speech Transcripts. CoRR. 2019;abs/1910.10238;Available from: http://arxiv.org/abs/1910.10238.
57. Zhang R, Zhang C. Dynamic Sentence Boundary Detection for Simultaneous Translation. In: Proceedings of the First Workshop on Automatic Simultaneous Translation. Seattle, Washington: Association for Computational Linguistics; 2020. p. 1-9;Available from: https://doi.org/10.18653/v1/2020.autosimtrans-1.1.
58. Luque FM. Atalaya at TASS 2019: Data Augmentation and Ro-

bust Embeddings for Sentiment Analysis. In: Cumbreras MÁG, Gonzalo J, Cámara EM, Martínez-Unanue R, Rosso P, Carrillo-de-Albornoz J, et al., editors. Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019. vol. 2421 of CEUR Workshop Proceedings. CEUR-WS.org; 2019. p. 561-70;Available from: https://ceur-ws.org/Vol-2421/TASS_paper_1.pfd.

59. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. ACL '02. USA: Association for Computational Linguistics; 2002. p. 311-318;Available from: https://doi.org/10.3115/1073083.1073135.

60. Google. Evaluating models automl translation documentation google cloud [Google.];Available from: https://cloud.google.com/translate/automl/docs/evaluate.

61. Jalili Sabet M, Dufter P, Yvon F, Schütze H. SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics; 2020. p. 1627-43;Available from: https://aclanthology.org/2020.findings-emnlp.147.

Tạp chí Phát triển Khoa học và Công nghệ – Kĩ thuật và Công nghệ 2023, 6(4):2099-2117
Open Access Full Text Article

Bài nghiên cứu

Tận dụng tăng cường dữ liệu tập trung theo câu và kiến trúc Transformer trong dịch tiếng Việt - tiếng Bana

Nguyễn Tấn Sang[1,2], Phạm Quốc Nguyên[1,2,*], Quản Thành Thơ[1,2]

TÓM TẮT
Bana là một dân tộc thiểu số ở Việt Nam với những di sản văn hóa phong phú, đồng thời sở hữu ngôn ngữ có ý nghĩa văn hóa và lịch sử vô cùng to lớn. Chính phủ chú trọng mạnh mẽ vào việc bảo tồn và thúc đẩy phát triển tiếng Bana bằng cách triển khai và khuyến khích sử dụng ở nhiều thế hệ. Các tiến bộ gần đây trong trí tuệ nhân tạo, như *Dịch Máy Nơ-ron* (NMT), đã đem lại những chuyển biến tốt trong dịch thuật nhờ cải thiện độ chính xác và sự trôi chảy. Điều này góp phần vào việc phục hồi tiếng Bana thông qua các nỗ lực trong giáo dục, giao tiếp và lưu trữ tài liệu. Cụ thể, NMT đóng vai trò then chốt trong việc cải thiện khả năng giao tiếp cho người Bana, làm cho việc tiếp cận thông tin và nội dung bên ngoài dễ dàng hơn. Tuy nhiên, việc dịch từ tiếng Việt sang tiếng Bana phải đối mặt với một thách thức lớn đó là hạn chế về tài nguyên, vì nguồn ngữ liệu đặc biệt ít ỏi trong tiếng Bana. Để giải quyết vấn đề này, chúng tôi áp dụng các kỹ thuật tiên tiến nhất trong NMT cùng với hai chiến lược tăng cường dữ liệu cho việc dịch tiếng Việt-Bana theo các lĩnh vực cụ thể. Trong phương pháp *tăng cường dữ liệu đa nhiệm*, các cặp câu mới được tạo ra thông qua các biến đổi. Những câu được tăng cường này sẽ được sử dụng trong quá trình huấn luyện. Mục tiêu của phương pháp này nhằm giới thiệu ngữ cảnh mới khi mà các tiền tố đơn lẻ không cung cấp đủ thông tin cho việc dự đoán chính xác từ tiếp theo. Phương pháp này cải thiện khả năng mã hóa và buộc bộ giải mã phải tập trung hơn vào các nguồn từ bộ mã hóa. Mặt khác, phương pháp *tăng cường dữ liệu giới hạn câu* dựa trên phương pháp tạo nhiễu ở cấp độ câu. Trong dịch máy nơ-ron, xử lý các lỗi liên quan đến cấu trúc ngữ pháp và phân hoạch câu rất khó khăn. Việc sử dụng *tăng cường dữ liệu giới hạn câu* có thể giúp cải thiện chất lượng dịch. Từ đó, một chiến lược tăng cường dữ liệu đơn giản được đề xuất. Quan trọng hơn, cả hai phương pháp đều linh hoạt và có thể được sử dụng với các mô hình dịch máy nơ-ron khác nhau. Ngoài ra, các phương pháp này không đòi hỏi các bước tiền xử lý dữ liệu phức tạp, hay huấn luyện các hệ thống bổ sung, hoặc thu thập dữ liệu bổ sung ngoài các tập dữ liệu song song hiện có.

Từ khoá: Tăng cường dữ liệu, dịch máy sử dụng tài nguyên thấp, dịch máy, tiếng Bana

[1]*Khoa Khoa Học và Kỹ Thuật Máy Tính, Trường Đại Học Bách Khoa Thành Phố Hồ Chí Minh, 268 Lý Thường Kiệt, Quận 10, Thành phố Hồ Chí Minh, Việt Nam*

[2]*Đại Học Quốc Gia Thành Phố Hồ Chí Minh, Phường Linh Trung, Thành Phố Thủ Đức, Thành Phố Hồ Chí Minh, Việt Nam*

Liên hệ

Phạm Quốc Nguyên, Khoa Khoa Học và Kỹ Thuật Máy Tính, Trường Đại Học Bách Khoa Thành Phố Hồ Chí Minh, 268 Lý Thường Kiệt, Quận 10, Thành phố Hồ Chí Minh, Việt Nam

Đại Học Quốc Gia Thành Phố Hồ Chí Minh, Phường Linh Trung, Thành Phố Thủ Đức, Thành Phố Hồ Chí Minh, Việt Nam

Email: quocnguyenkh@gmail.com

Lịch sử
- Ngày nhận: 15-10-2023
- Ngày chấp nhận: 25-3-2024
- Ngày đăng: 13-5-2024

DOI : https://doi.org/10.32508/stdjet.v6i4.1284

Trích dẫn bài báo này: Sang N T, Nguyên P Q, Thơ Q T. **Tận dụng tăng cường dữ liệu tập trung theo câu và kiến trúc Transformer trong dịch tiếng Việt - tiếng Bana**. *Sci. Tech. Dev. J. - Eng. Tech.* 2023, 6(4):2099-2117.