Department of Information and Communication Technology

Natural Language Processing

Vietnamese Text Correction and Spell Checking

Group 5
22BI13472 - Nguyễn Bá Vinh
22BI13220 - Nguyễn Minh Khôi
22BI13227 - Trần Trung Kiên
22BI13462 - Chu Hoàng Việt
22BI13351 - Nguyễn Ngọc Nhi
22BI13352 - Vũ Hoàng Mai Nhi

Academic year: 2022 - 2025
Hanoi, March 2025

Contents

1 Introduction

This section provides an overview of the challenges in Vietnamese error correction and the goals of the proposed model. It also introduces the structure of the dataset and the output from the model.

1.1 Overview

The Vietnamese language presents unique challenges for error correction due to its complex spelling, grammar, word order, and tonal system. These intricacies make it difficult to automatically detect and correct errors in text. Despite these challenges, there is a growing need for efficient tools to improve the quality of Vietnamese text, particularly in applications such as text processing, machine translation, and content generation.

In response to this challenge, we aim to build a Vietnamese error correction model designed to enhance the quality of Vietnamese text. Our model will focus on fixing common issues such as spelling mistakes, grammatical errors, and typographical errors. By addressing these problems, the model will contribute to the improvement of automated Vietnamese language processing and enable more accurate and reliable text analysis in various applications.

1.2 Input

The input comprises Vietnamese text containing various typographical and spelling errors. Below are two representative examples of erroneous sentences from the dataset:

- `"Côn viec kin doanh thì rất kho khan nên toi quyết dinh chuyển sang nghề khac."`
- `"Toi dang là sinh diên nam hai ở truong đạ hoc khoa jọc tự nhiên, trogn năm ke tiep toi sẽ chọn chuyen nganh về trí tue nhana tạo."`

These examples illustrate the type of noisy input provided to the NLP model, which is designed to preprocess the text and correct typographical and spelling errors.

1.3 Output

The model successfully corrected spelling and typographical mistakes in the input text, producing the following sentences:

- `"Công việc kinh doanh thì rất khó khăn nên tôi quyết định chuyển sang nghề khác."`
- `"Tôi đang là sinh viên năm hai ở trường đại học khoa học tự nhiên, trong năm kế tiếp tôi sẽ chọn chuyên ngành về trí tuệ nhân tạo."`

2 Dataset

In this section, we discuss the technical backbone of the system, and how data is gathered, processed, and used to power our system.

2.1 Purpose

The primary purpose of this dataset is to support the development and training of machine learning models that can automatically detect and correct errors in Vietnamese text. These errors could range from common spelling mistakes to more complex grammatical issues, including but not limited to missing diacritical marks, incorrect word order, and verb conjugation mistakes. By providing a large collection of erroneous text paired with corrected text, the dataset enables the training of models that can not only fix individual errors but also understand context to correct grammatical mistakes.

2.2 Training Set

The training dataset was from **huggingface/bmd1905/error-correction-vi**, which consists of nearly 359,993 rows of Vietnamese sentences, each containing an erroneous sentence (`error_text`) and its corresponding corrected version (`correct_text`).

- **error_text:** This column contains sentences with common typographical errors, misspelled words, missing accents, incorrect diacritics, and other grammatical errors that often occur in Vietnamese writing. These errors simulate real-world mistakes made by writers, either due to quick typing, lack of familiarity with proper Vietnamese grammar and spelling conventions, or informal writing styles. Common issues in error_text include missing spaces, incorrect word order, phonetic mistakes, and wrong use of uppercase letters. The dataset helps capture these real-world writing challenges and provides a basis for error correction tasks.
- **correct_text:** This column contains the corrected sentences, which have been edited to remove errors and align with proper Vietnamese grammar and spelling. These corrections are manually curated to ensure that the text is grammatically correct and uses standard Vietnamese orthography.

The dataset is well-suited for training models in various Vietnamese language tasks, including spell-checking, grammar correction, and text normalization.

2.3 Testing Set

For evaluation purposes, the test set is composed of 10,000 rows, which are selected from the total 359,993 rows in the **huggingface/bmd1905/error-correction-v2** dataset. The test set maintains the same format as the training set, consisting of two columns which are error_text and correct_text.

The dataset, with a total size of 131 MB, provides a substantial amount of data for model training and testing. The test set was specifically chosen to ensure a diverse representation of various types of errors and corrections, allowing for a thorough evaluation of the model's performance in handling real-world text errors.

2.4 Dataset Example

Below table is an example of a row in the dataset we used:

Column Name	Example
error_text	"A Hùnig hả ra đây đang uống biAa cùng Mạnh Hà Alô Dũngà bảo đến ngay mà qgiờ này vẫn chưa thẤy đâu tao say rồi"
correct_text	"A Hùng hả, ra đây, đang uống bia cùng Mạnh, Hà... Alô Dũng à, bảo đến ngay mà giờ này vẫn chưa thấy đâu, tao say rồi..."

3 Model

This section describes the architecture, the pretraining data and the optimization setup, that we use for our model - BARTpho.

3.1 Architecture

In this project, we use Bidirectional and Auto-Regressive Transformer model. BART is a powerful sequence-to-sequence (seq2seq) model that integrates both bidirectional and autoregressive components, making it highly effective for a wide range of natural language processing tasks.

The model, based on the encoder-decoder framework, is pre-trained as a denoising autoencoder and fine-tuned for tasks like text summarization, machine translation, and text generation. By combining the bidirectional encoder of BERT (Bidirectional Encoder Representations from Transformers) with the autoregressive decoder of GPT (Generative Pre-trained Transformer), BART leverages the strengths of both architectures. As a denoising autoencoder, BART is trained by corrupting the input text—introducing noise through methods like sentence permutation, text infilling, or token masking—and then learning to reconstruct the original version. The encoder part of the model is responsible for analyzing and restructuring the input sentences by capturing contextual relationships between words. It processes the input data and creates a representation that highlights the underlying structure. On the other hand, the decoder's role is to restore or predict the original sentence by generating the correct output based on the encoded information, ensuring that the final sentence is grammatically accurate and matches the intended meaning.

The attention mechanism employed by both the encoder and decoder enables the model to focus on relevant parts of the input when correcting errors. This process

enables the model to understand contextual relationships in text, making it highly effective at correcting spelling, typographical, and grammatical errors in Vietnamese.

3.2 Data Flow

This section explains how data moves during the text correction process, from raw input text with errors to the generation of a corrected output sentence.

- **Input**: The user provides raw Vietnamese text that may contain typographical or grammatical errors.
  Example: "bàn thắng duy nhẤt của trận đấu được ghì ở phút 85, do cWông củn tiền đạz MAnit Minh Hải"

- **Tokenization**: The input text is processed by the tokenizer associated with the PhoBART model. This tokenizer splits the input into tokens using a syllable-based tokenization approach, ensuring compatibility with the nuances of the Vietnamese language.

- **Embedding Layer**: Each token generated by the tokenizer is converted into a high-dimensional numerical vector, known as an embedding. These embeddings capture the semantic and syntactic meaning of each token within the context of the Vietnamese language.

- **Encoding**: The token embeddings are fed into the PhoBART Encoder, a key component of the BART architecture. The encoder uses self-attention mechanisms and feedforward neural layers to process the sequence of embeddings. This step extracts contextual relationships between tokens, enabling the model to understand the dependencies and interactions among words. The encoder outputs a set of contextual embeddings, which represent the entire input sequence.

- **Decoding**: The contextual embeddings from the encoder are passed to the PhoBART Decoder, which generates the corrected text. The decoder operates in an auto-regressive manner, predicting one token at a time. Each token prediction is informed by the contextual embeddings and the previously generated tokens, ensuring that the output is grammatically correct, fluent, and semantically accurate.

- **Output**: The predicted tokens are combined to form the final corrected sentence.
  Example: "Bàn thắng duy nhất của trận đấu được ghi ở phút 85, do công của tiền đạo Manit. Minh Hải."

3.3 Work Flow

This section describes the workflow of the model, focusing on its fine-tuning and evaluation process for Vietnamese text correction. It begins with fine-tuning PhoBART, on the training dataset containing pairs of erroneous and corrected sentences. The fine-tuned model's performance is then evaluated using a subset of the dataset, where its generated corrections are compared against ground-truth references using the BLEU score.

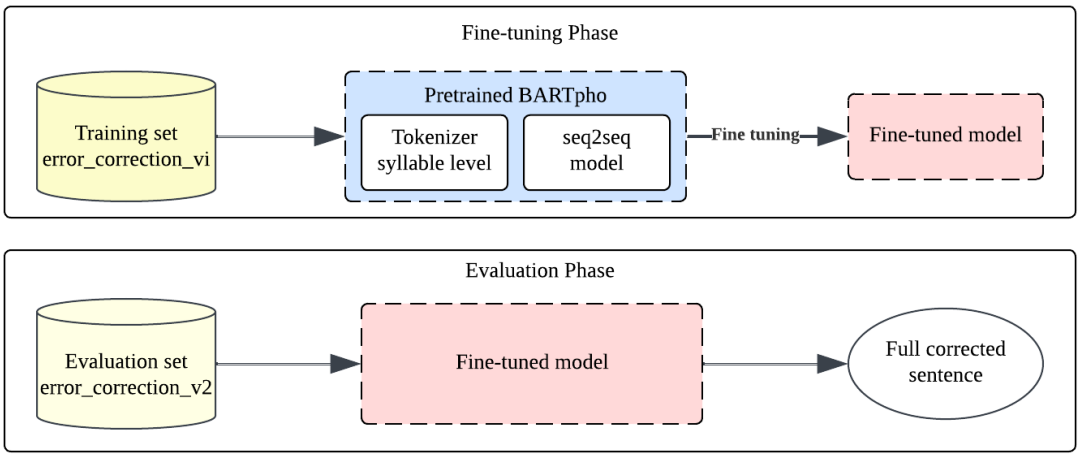

Figure 1: Work Flow

3.3.1 Fine-tuning PhoBART for Vietnamese Text Correction

The foundation of this project is PhoBART (vinai/bartpho-syllable), a pre-trained sequence-to-sequence transformer model based on the BART architecture. This model was pre-trained on a large corpus of Vietnamese text to understand the structure, syntax, and semantics of the language.

PhoBART was fine-tuned specifically for the task of Vietnamese text error correction using the vi-error-correction dataset, which consists of pairs of erroneous input sentences and their corrected counterparts. After this process, we obtain an efficient model for the task of text error-correction.

3.3.2 Evaluating the Fine-tuned Model

After the fine-tuning phase, its performance was evaluated using a subset of 10,000 pairs of erroneous and corrected sentences randomly sampled from the vi-error-correction dataset-v2, which contains a total of 359,993 pairs. Each pair consists of an erroneous input sentence and its corresponding corrected sentence.

The model was tested by providing the erroneous sentences as input and generating corrected outputs. These outputs were then compared against the ground-truth corrected sentences using the BLEU score as the evaluation metric. The BLEU score measures the similarity between the generated corrections and the reference sentences, highlighting the fine-tuned model's effectiveness in addressing real-world Vietnamese text correction tasks.

4 Result

This section presents the evaluation results of the fine-tuned PhoBART model on the Vietnamese text correction task. The model was tested using 100 samples of the total dataset from vi-error-correction-v2, which corresponds to a representative subset of the 359,993 sentence pairs. The evaluation metrics used include the BLEU score, which measures the similarity between the generated corrections and the reference sentences, and loss entropy, which indicates how well the model predicts the corrected outputs.
The results are summarized in the table below:

Metric	**Score**
BLEU Score	91.33
Loss Entropy	0.11

Table 1: Evaluation metrics for the fine-tuned PhoBART model on 1% of the vi-error-correction-v2 dataset.

These metrics highlight the effectiveness of the fine-tuned model in generating accurate corrections for Vietnamese text errors. The BLEU score reflects the closeness of the model's outputs to the ground-truth corrections, while the loss entropy demonstrates the model's predictive accuracy during testing.

Below is an example of the testing process with a sentence in the dataset vi-error-correction-v2:

- Input: "Sau đó bài hátđược nhiều người biết kđến yơn với sự thể hiện"
- Output: "Sau đó bài hát được nhiều người biết đến hơn với sự thể hiện."

5 Future Work

There are several ways to improve the Vietnamese text correction model. First, enhancing the accuracy of error correction by adding more contextual understanding could help the model handle more complex grammar and errors. This could involve integrating syntax or semantics to improve how the model understands sentence structure.

Expanding the dataset is another important step. Training the model on more diverse data, such as social media text, informal writing, or specific areas like medical or legal texts, would make it more versatile. Including different Vietnamese dialects would also help the model work better across the country.

Another key area is fine-tuning the model for specific domains, such as legal, medical, or technical fields. This would improve its ability to handle industry-specific language and terminology.

Real-time error correction is another possible improvement. Developing a version of the model that works instantly as users type in text editors or messaging apps would enhance user experience.

Incorporating user feedback into the model could allow it to improve over time. By learning from corrections made by users, the model could become more personalized and accurate.

Finally, testing the model on more datasets and comparing it to other models would help identify areas for improvement. Experimenting with different pre-training strategies would also make the model more robust to various types of errors.

In conclusion, future improvements to the model could make it more accurate, versatile, and user-friendly, with broader applications in real-time systems, specific industries, and multiple languages.

References

Bui, H. (2025). error-correction-vi. [online] Huggingface.co. Available at: https://huggingface.co/datasets/bmd1905/error-correction-vi [Accessed 27 Mar. 2025]

Bui, H. (2025). vi-error-correction-v2. [online] Huggingface.co. Available at: https://huggingface.co/datasets/bmd1905/vi-error-correction-v2 [Accessed 27 Mar. 2025]

Tran, N.L., Le, D.M. and Nguyen, D.Q. (2022). BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese. arXiv:2109.09701 [cs]. [online] Available at: https://arxiv.org/abs/2109.09701

GitHub - vinhngba2704Group-5_NLP. [online] GitHub. Available at: https://github.com/vinhngba2704/Group-5_NLP.git [Accessed 27 Mar. 2025].