Research on Transformer model and application of Vietnamese diacritic restoration

Nong Hong Lanh

Abstract: This study investigates the application of the Transformer model to the task of Vietnamese diacritic restoration, a critical challenge in natural language processing due to the omission of tonal marks in informal digital text. While traditional methods often rely on recurrent neural networks or statistical approaches, the Transformer architecture offers superior capabilities in capturing long-range contextual dependencies through its self-attention mechanism. We propose a sequence-to-sequence model based on the Transformer, fine-tuned on a large corpus of Vietnamese text with and without diacritics. The model is evaluated on multiple benchmark datasets, demonstrating significant improvements in accuracy and fluency compared to baseline models such as BiLSTM-CRF. Experimental results show that the Transformer effectively resolves ambiguities arising from homographic words and complex tonal patterns. This work highlights the potential of attention-based architectures for low-resource language tasks and provides a practical solution for enhancing Vietnamese text processing in applications like machine translation, speech recognition, and text normalization.

Keywords: Transformer, Vietnamese diacritic restoration, natural language processing, diacritic restoration