HYBRID METHOD COMBINING HIERARCHICAL TRANSFORMER ENCODERS AND SEQUENCE-TO-SEQUENCE FOR VIETNAMESE SPELLING CORRECTION

Trần Văn Hoàng - 230101043

Summary

- Course: CS2205.CH181
- Link Github: https://github.com/hoangftran/CS2205.CH181
- Full name: Trần Văn Hoàng - 230101043

Introduction

Given a text, **identify** misspelling or misplaced words and **suggest** suitable replacements within the **context** of a sentence.

The importance of Vietnamese Spell Correction:

- Spell correction is critical in Vietnamese NLP due to complex tonal and diacritical systems
- Misspellings can significantly impact downstream tasks
- Current approaches have limitations in handling varying input/output lengths and diverse error types

Related works

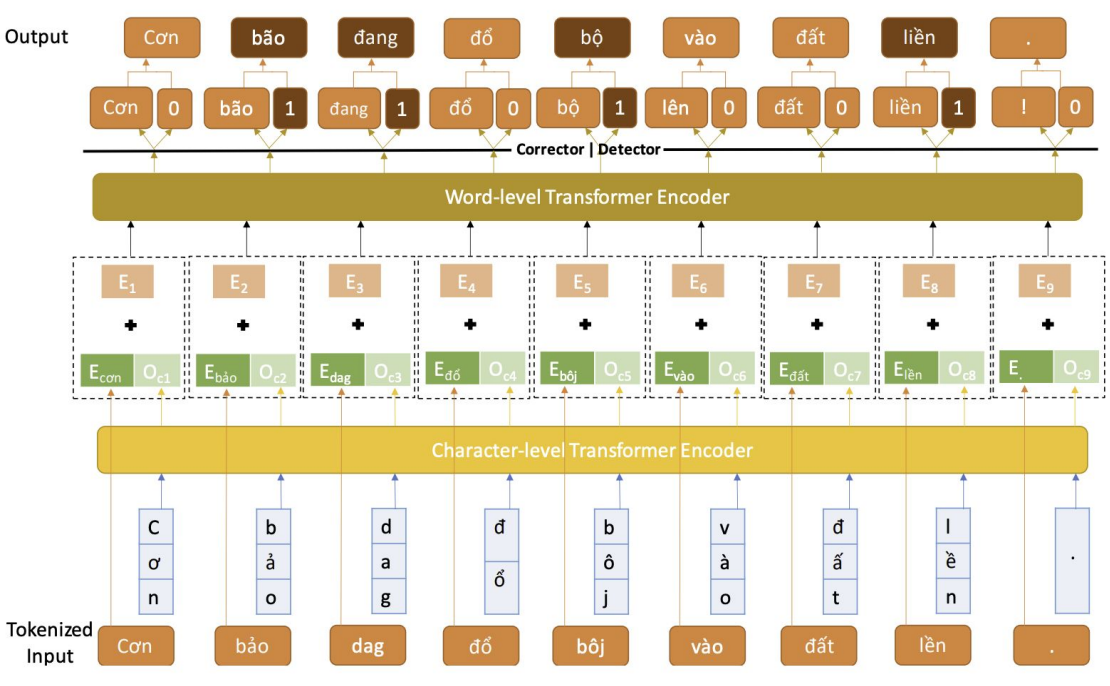

Fig 1. Hierarchical Transformer Encoders [2]

Related works

VSEC: Transformer-based Model for Vietnamese Spelling Correction [3]

- Treating Vietnamese spelling error correction as a machine translation problem
- Use the seq2seq architecture based on Transformer as baseline

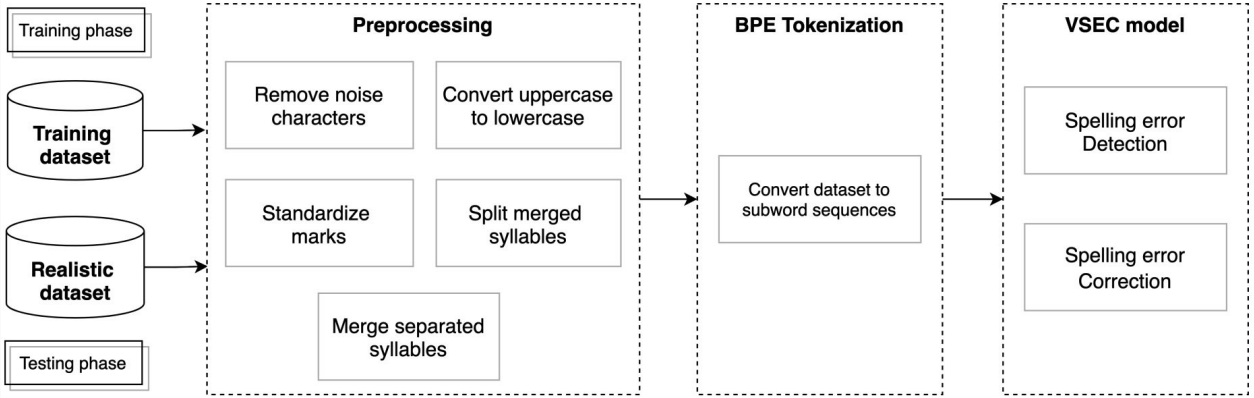

Fig 2. VESC pipeline

Objectives

Main objectives

1. Develop a novel hybrid architecture combining Hierarchical Transformer Encoders and Transformer-based models
2. Create a comprehensive Vietnamese spell correction dataset reflecting real-world error types
3. Achieve state-of-the-art performance in Vietnamese spell correction

Methodology

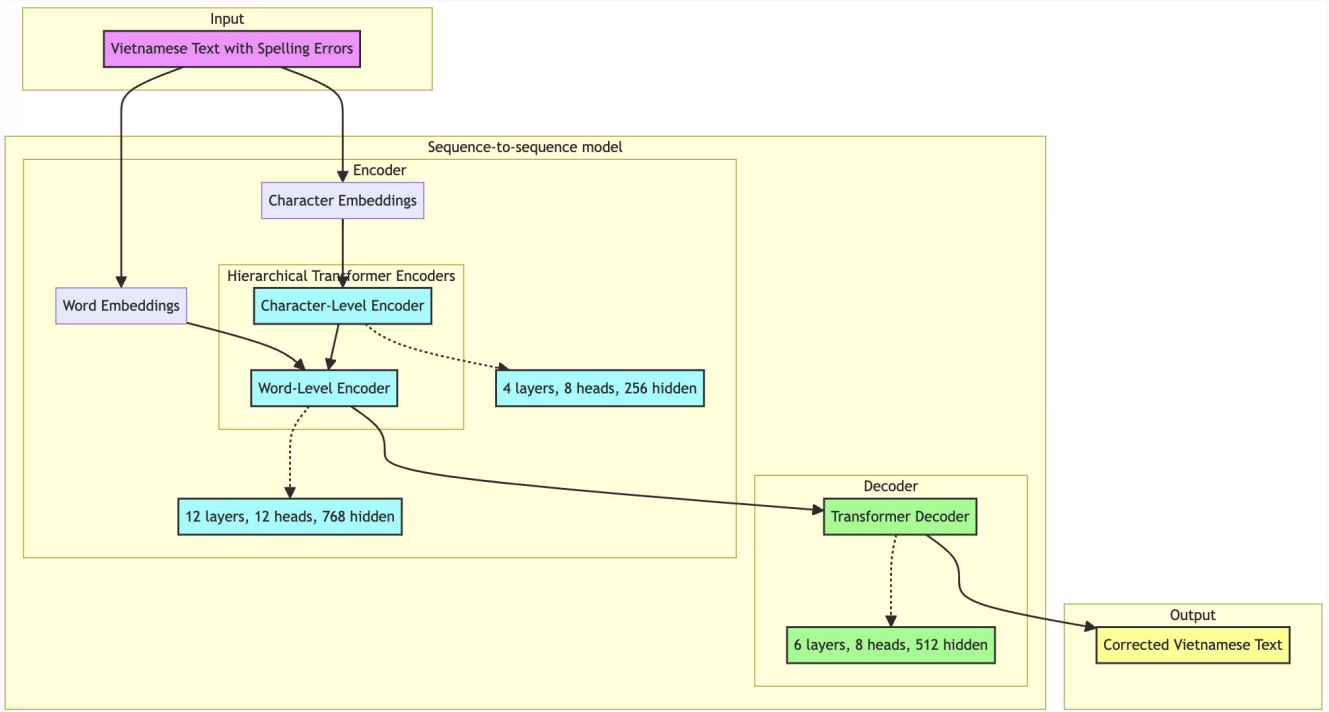

Fig 3. Proposed architecture

Methodology

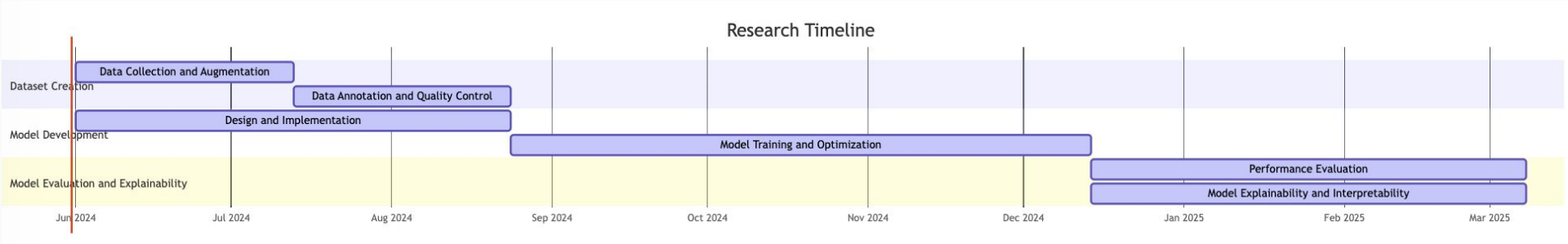

Fig 4. Proposed timeline

Expected result

Dataset and Model

1. A Comprehensive Vietnamese Spell Correction Dataset
2. A State-of-the-Art Hybrid Vietnamese Spell Correction Model

Robustness, Insights, and Efficiency

3. Improved Robustness and Generalizability
4. Insights into Model Behavior and Error Patterns
5. Efficiency and Scalability

References

[1]. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention Is All You Need. arXiv: 2111.00640

[2]. Hieu Tran, Cuong V. Dinh, Long Phan, Son T. Nguyen. Hierarchical Transformer Encoders for Vietnamese Spelling Correction. The 34th International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems, 2021

[3]. Dinh-Truong Do, Ha Thanh Nguyen, Thang Ngoc Bui, Dinh Hieu Vo. VSEC: Transformer-based Model for Vietnamese Spelling Correction. arXiv preprint arXiv: 2111.00640, 2021

[4]. Thanh-Nhi Nguyen, Thanh-Phong Le, Kiet Van Nguyen. ViLexNorm: A Lexical Normalization Corpus for Vietnamese Social Media Text. The 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2024