A Transformer-Based Approach for Thai Handwritten Text Recognition with Subword Tokenization

Tichaiyut Thanapasee, Pakaidown Palanamai, Pongsapat Sangkaew, Warit Sirikosityanggoon
Advisor: Assistance Professor Dr.Papangkorn Inkeaw

FACULTY OF SCIENCE
CHIANG MAI UNIVERSITY

DSRC
Data Science Research Center
Faculty of Science, Chiang Mai University

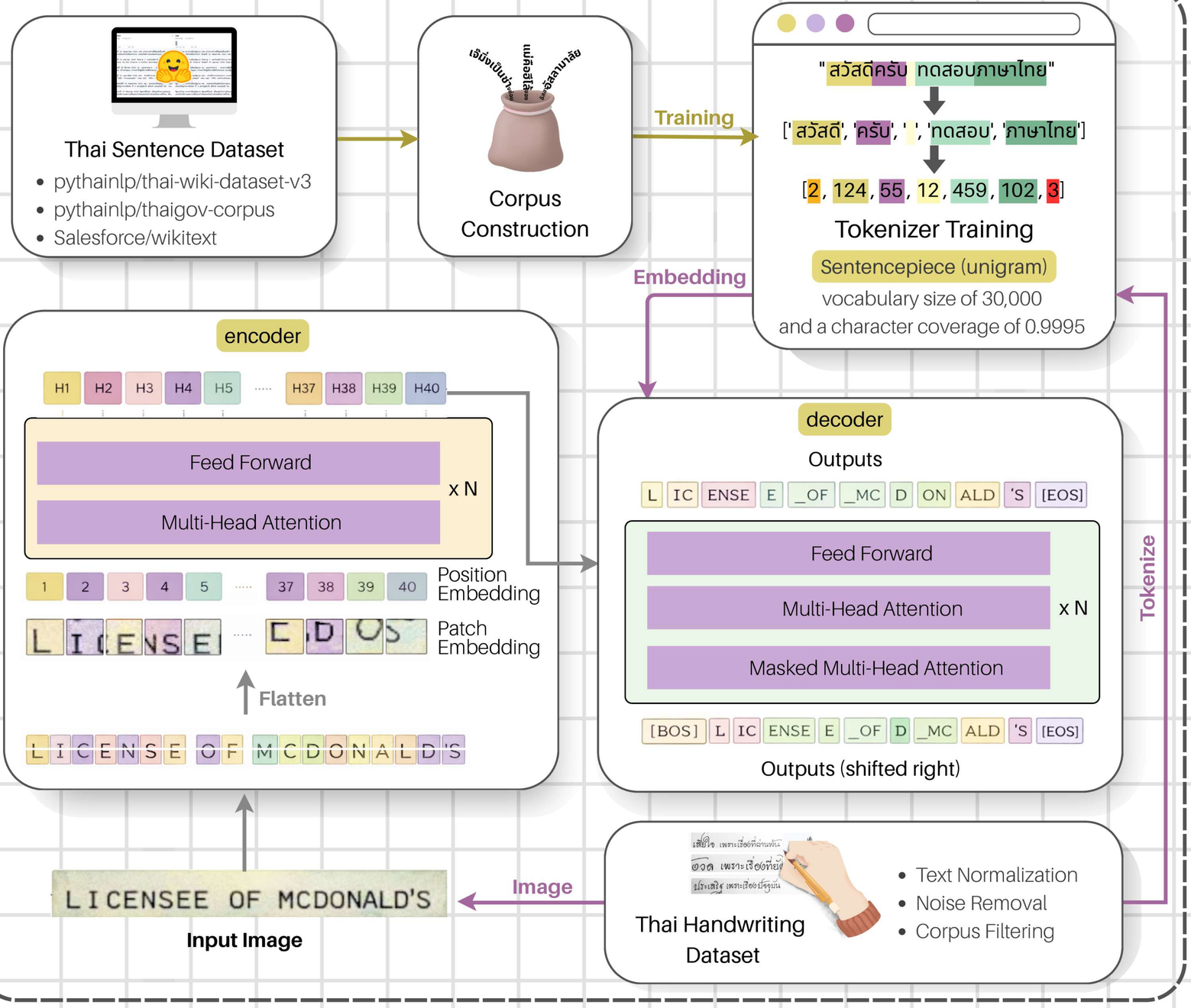
Abstract

Thai handwritten optical character recognition (OCR) remains a challenging task due to variations in writing styles, character connectivity, and the complexity of the Thai writing system. Transformer-based Optical Character Recognition (TrOCR) is a state-of-the-art deep learning architecture that enables end-to-end text recognition without explicit character segmentation. In this study, we adopted TrOCR to convert Thai handwritten images into text. A SentencePiece Unigram tokenizer with a vocabulary size of 30,000 was employed to provide effective subword representations, enabling the model to better handle out-of-vocabulary words and misspellings. A model pretrained on English handwritten data was fine-tuned using Thai handwritten datasets to leverage transfer learning. The proposed approach was evaluated on an independent Thai handwritten test set. It achieved a character error rate (CER) of 0.4244, character accuracy of 0.7416, and sequence match accuracy of 0.6614. The experimental results demonstrate that the proposed model outperforms baseline systems, including OpenThaiGPT, EasyOCR, and Tesseract. Furthermore, the use of subword tokenization significantly contributed to performance improvement.

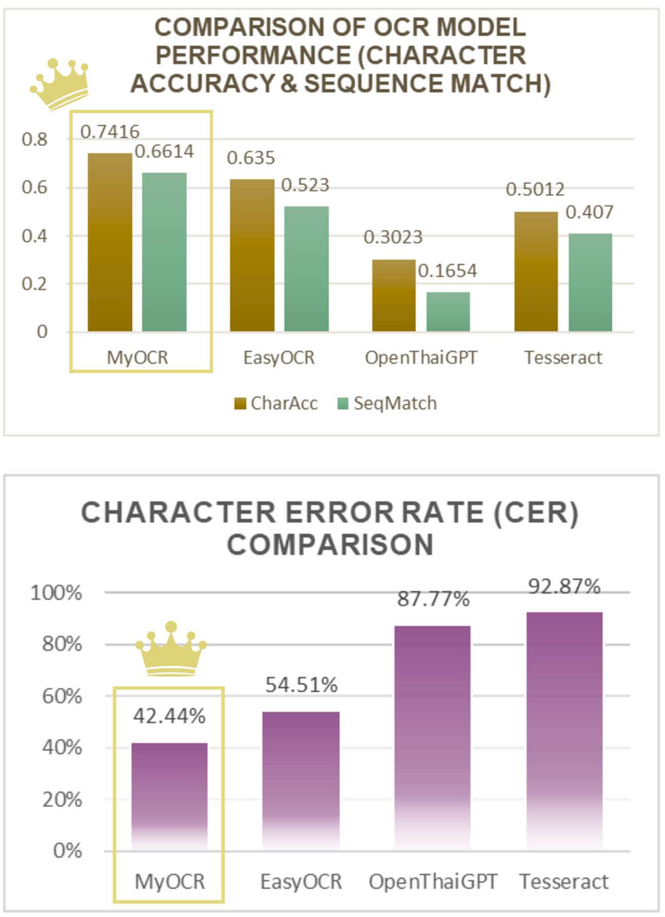
Introduction & Objective

- Handwritten text recognition is essential for document digitization and knowledge preservation. However, Thai handwriting is particularly challenging due to its complex character structure and writing patterns.
- Although Transformer-based models have shown strong performance in vision and language tasks, their application to Thai handwritten text recognition remains limited and often does not fully address Thai linguistic characteristics.
- This study proposes a Transformer-based approach with a token representation strategy tailored to Thai, aiming to improve recognition performance and support low-resource language applications.

Methodology



Result



Evaluation Metrics

Character Error Rate (CER)

$$CER = \frac{S + D + I}{N}$$

S = substitutions
D = deletions
I = insertions
N = total characters in the ground truth

Evaluation Metrics

Character Accuracy

$$CharAcc = \frac{|P \cap G|}{|G|}$$

P = the sequence of characters predicted
G = the sequence of characters in the ground-truth

Evaluation Metrics

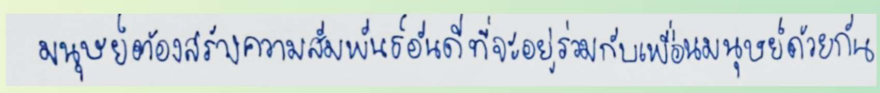
LCS-based Sequence Match

$$SeqMatch = \frac{LCS(P,G)}{|G|}$$

LCS(P, G) = denotes the length of the longest common subsequence
|G| = the length of the ground-truth text

Output

Input Image



Predicted Text

มนุษย์ต้องความสัมพันธ์อันดีที่จะอยู่กับเพื่อนมนุษย์ด้วยกัน

Conclusion

Overall, the findings confirm that combining Transformer-based architectures with subword tokenization is an effective strategy for Thai handwritten OCR. The proposed approach demonstrates that a Transformer encoder–decoder framework can accurately recognize Thai handwritten text while maintaining robustness to linguistic complexity. Beyond its performance, this work provides a foundation for extending handwritten text recognition to other low-resource languages. It also supports the digital preservation of linguistic and cultural documents in Thailand by enabling the conversion of handwritten materials into accessible digital text.

References

Li, M., Lv, T., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., & Wei, F. (2021). TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. arXiv preprint arXiv:2109.10282. [https://arxiv.org/abs/2109.10282]
OpenThaiGPT. (2024). Thai TrOCR: Thai Handwriting Recognition Model. Hugging Face. [https://huggingface.co/openthaigpt/thai-trocr]