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

Author(s) :

1. Tichaiyut Thanapasee	**Student ID :**	650510707
2. Pakaidown Palanamai	**Student ID :**	650510731
3. Pongsapat Saengkaew	**Student ID :**	650510732
4. Warit Sirikosityanggoon	**Student ID :**	650510735

Major : Data Science

Advisor(s) : **1.** Assistant Professor Dr. PAPANGKORN INKEAW

Type of presentation* ***(choose 1)*** **:**

- ☐ **Oral Presentation** (เฉพาะ ตัวแทนนศ.ที่สาขาเลือกให้นำเสนอแบบบรรยาย)
- ☑ **Poster** (กรณี นำเสนอผลงานปัญหาพิเศษ/การค้นคว้าอิสระ)
- ☐ **Cooperative Education** (กรณี นำเสนอผลงานสหกิจศึกษา)

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.

**Type of presentation must be matched with an option you choosing on student upload system.*

***The abstract can be more than one page and must be approved by project advisor before upload.*

Title name guide.

ADVISOR title name / แปลไทย	
Professor Dr.	ศาสตราจารย์ ดร.
Professor	ศาสตราจารย์
Associate Professor Dr.	รองศาสตราจารย์ ดร.
Associate Professor	รองศาสตราจารย์
Assistant Professor Dr.	ผู้ช่วยศาสตราจารย์ ดร.
Assistant Professor	ผู้ช่วยศาสตราจารย์
Dr.	ดร.
Lecturer	อาจารย์
Mrs.	นาง
Ms.	นางสาว
Mr.	นาย

Major name guide.

SCIENCE MAJOR name / แปล	
Biology	ชีววิทยา
Microbiology	จุลชีววิทยา
Zoology	สัตววิทยา
Biochemistry and Biochemical Technology *or* Biochemistry and Biochemical Innovation	ชีวเคมีและชีวเคมีเทคโนโลยี *หรือ* ชีวเคมีและชีวเคมีนวัตกรรม
Chemistry	เคมี
Industrial Chemistry	เคมีอุตสาหกรรม
Materials Science	วัสดุศาสตร์
Physics	ฟิสิกส์
Computer Science	วิทยาการคอมพิวเตอร์
Data Science	วิทยาการข้อมูล
Mathematics	คณิตศาสตร์
Statistics	สถิติ
Gemology	อัญมณีวิทยา
Geology	ธรณีวิทยา
Environmental Science	วิทยาศาสตร์สิ่งแวดล้อม

**Type of presentation must be matched with an option you choosing on student upload system.*

***The abstract can be more than one page and must be approved by project advisor before upload.*