B.Comp. Dissertation

Benchmarking and Improving OCR System for Southeast Asian Languages

By

Qiu Jiasheng, Jason

Department of Computer Science

School of Computing

National University of Singapore

2024/2025

B.Comp. Dissertation

Benchmarking and Improving OCR System for Southeast Asian Languages

By

Qiu Jiasheng, Jason

Department of Computer Science

School of Computing

National University of Singapore

2024/2025

Project ID: H0792230

Supervisor: A/P Min-Yen Kan

Advisor: Tongyao Zhu

Abstract

While Optical Character Recognition (OCR) has been widely studied for high-resource languages such as English and Chinese, the efficacy and limitations of OCR models on Southeast Asian (SEA) languages remain largely unexplored. This study aims to bridge this gap by assessing and improving the performance of OCR technologies on SEA languages. To achieve this objective, we propose a reusable pipeline to gather SEA-language text from Wikipedia and benchmark popular OCR tools.

Subject Descriptors:

I.2.7 Natural Language Processing

Keywords:

Optical Character Recognition, Southeast Asian Languages

Implementation Software and Hardware:

Python, Tesseract, EasyOCR

Acknowledgements

I would like to thank my supervisor, A/P Kan Min-Yen, and my advisor, Tongyao Zhu, for their invaluable guidance and mentorship. Their encouragement and constructive guidance have been a significant source of inspiration throughout the project.

List of Figures

List of Tables

Table of Contents

Introduction

Current research in Natural Language Processing (NLP) is heavily concentrated on 20 of the 7,000 languages in the world (Magueresse et al., 2020). In particular, Southeast Asia (SEA) is home to over 1,000 languages but remains a relatively under-researched region in NLP (Aji et al., 2023). Similar to most low-resource languages, a major challenge in developing NLP systems for SEA languages is the limited availability of datasets for the region's languages. Although many scanned documents and books in these low-resource languages are available online, the text within these files remains inaccessible due to formats like images and PDFs.

A solution to this problem is to use Optical Character Recognition (OCR) to extract the textual data. OCR is the process of identifying and converting text in an image into a computer-friendly text format. By extracting the text from these scanned documents, OCR can generate valuable datasets for low-resource languages. The created datasets can then be used for downstream NLP tasks, such as machine translation, training large language models, and named-entity recognition (Agarwal & Anastasopoulos, 2024; Ignat et al., 2022). Therefore, studying OCR performance on SEA languages is crucial to accelerating NLP research in the region.

While OCR has been widely studied for high-resource languages such as English and Chinese, the efficacy and limitations of OCR models on SEA languages remain largely unexplored. To address this gap, we propose a reusable pipeline to collect textual data in low-resource SEA languages from Wikipedia and benchmark popular open-source OCR tools on the collected data. The primary objective is to

benchmark and improve the performance of OCR technologies on SEA languages, thereby contributing to the advancement of NLP applications in this linguistically diverse region. Specifically, this project seeks to answer the following research questions (RQs):

- **RQ1.** How do popular OCR tools perform on SEA scripts?
- **RQ2.** What specific linguistic and script-related challenges affect OCR accuracy on SEA languages?
- **RQ3.** What techniques and recommendations can enhance OCR accuracy on SEA languages?

Conclusion