Informatics Institute of Technology

In Collaboration With

University of Westminster, UK

University of Westminster, Coat of Arms

SWA BHASHA 2.0: SOLVING AMBIGUITIES IN ROMANIZED SINHALA TO SINHALA TRANSLITERATION USING NEURAL MACHINE TRANSLATION

A Dissertation by

Mr. Sachithya Rivisara Dharmasiri

w1761185 / 2019099

Supervised by

Mr. Deshan Sumanathilaka

May 2023

Submitted in partial fulfilment of the requirements for the BEng (Hons) Software Engineering at the University of Westminster.

ABSTRACT

The widespread adoption of social media and instant messaging has made it essential to communicate in one's native tongue. Romanized Sinhala and native Sinhala are both frequently used in Sinhala but attempts to use machine transliteration to transliterate Romanized Sinhala to native Sinhala can result in inaccuracies. This is due to the informal text shorthand known as "Singlish-based shorthand words" Rule-based transliteration systems may not be compatible with the ad hoc transliterations used in Singlish. To address this issue, a **Novel hybrid approach combining rule-based machine translation and neural machine translation** has been proposed. Combining the advantages of rule-based algorithms and neural machine translation, the proposed transliterator has the potential to considerably enhance reverse transliteration and improve communication in native Sinhala by combining the strengths of both approaches.

Using multiple metrics, the implemented neural machine translation model has been evaluated. The **BLEU score of 0.84** indicates that Sinhala transliterations generated from Romanized Sinhala text are accurate. In addition, the WER score of 0.16 demonstrates the model's ability to transcribe Sinhala text from its Romanized form accurately. The model accurately predicted the Sinhala transliteration in **84%** of the test cases, as indicated by the accuracy score of 0.84. Precision and recall scores of 0.861 and 0.862, respectively, indicate that the model accurately identified Sinhala words and their transliterations. The F1-score of 0.723 indicates that the model is well-balanced regarding precision and recall. In every test case, the ROUGE-L score of 1.00 indicates that the model obtained perfect overlap between the generated and reference Sinhala transliterations.

Due to the rule-based approach's low BLEU score, it was utilized as a part of the suggestion component rather than the primary transliteration component. Even though the preliminary test results are promising, additional testing and refinement are necessary to improve the overall performance of the machine translation models. The hybrid approach proposed has the potential to considerably enhance communication in native Sinhala and reverse transliteration.

Keywords: Romanized Sinhala, Singlish, machine transliteration, rule-based machine translation, neural machine translation, suggestion algorithm, reverse transliteration, BLEU score.