INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

Translate Sinhala Pseudocode to C#

A Natural Language Processing Approach

A Dissertation by

Mr. Pubudu Senarathne

w1898859 / 20211185

Supervised by

Mr. Yohan Pandigama

September 2023

Submitted in partial fulfilment of the requirements for the MSc in Advanced Software Engineering degree at the University of Westminster.

ABSTRACT

In the realm of software development and algorithm deployment, the contribution of automatic code generation is indispensable. In Sri Lanka, Sinhala is widely spoken and used in various domains, including education and research. However, the lack of tools for translating Sinhala pseudocode into programming languages like C# has been a significant obstacle in fostering local talent and promoting software development in the Sinhala language.

This research contributes to the fields of natural language processing, machine translation by utilizing graph theory, compiler theory and code generation techniques. The developed application empowers Sinhala-speaking developers and students to express their ideas and algorithms in their native language while seamlessly generating executable C# code. This bridges the language barrier for them in using ACG tools while linking Sinhala and C# programming language as well.

Evaluations were critical to ensuring the reliability of this translation. Using the BLEU metric, a score of 0.9056 was achieved for the translation accuracy across a set of sample translations. Furthermore, the heuristic engine introduced for suggesting improvements in the event of keyword mistypes demonstrated impressive figures with an accuracy of 96.31%, precision of 96.31%, recall of 100.00%, an F1 score of 98.12%, and a Top-3 Accuracy of 99.72%. Each translated code was also compiled and executed manually, boasting a success rate of 85%. These results validate the system's efficacy and precision.

This study opens new possibilities for software development in the Sinhala language and promotes the growth of the local programming community.

Keywords: Automatic Code Generation, Graph theory, compiler theory, Pseudocode, C#, Sinhala, Natural language processing

Subject Descriptors:

• Theory of computation → Compiler theory → Code generation → Automatic code generation.