NEURAL MACHINE TRANSLATION APPROACH FOR SINGLISH TO ENGLISH TRANSLATION

H.G.Dinidu Sandaruwan

(189396L)

Degree of MSc in Artificial Intelligence

Department of Computational Mathematics

University of Moratuwa
Sri Lanka

February 2021

NEURAL MACHINE TRANSLATION APPROACH FOR SINGLISH TO ENGLISH TRANSLATION

H.G.Dinidu Sandaruwan
(189396L)

Thesis submitted in partial fulfilment of the requirement of the degree of MSc in Artificial Intelligence

Department of Computational Mathematics

University of Moratuwa
Sri Lanka

February 2021

Declaration

I declare that this is my own research dissertation and this does not incorporate without acknowledgement any material previously published for a Degree or a Diploma in any other university or institute of higher education and to the best of my knowledge, awareness and belief it does not include any content that previously published or composed by another person except where the acknowledgement is made in the text. Here, I also agree to photocopy and interlibrary loan of my dissertation (if accepted), and provide its title and abstract to external organizations.

Name of the Student	Signature:
H.G.Dinidu Sandaruwan	Date: 08th February 2021
Supervised by	Signature
Dr. Subha Fernando	Date: 08th February 2021
Supervised by	Signature
Dr. Sagara Sumathipala	Date: 08th February 2021

Abstract

This dissertation is for a research that aimed at proposing a language model to translate texts written in Singlish to English. Singlish is an alternative writing system for Sinhala language that uses Latin scripts (English Alphabet) instead of using native Sinhala alphabet. This had been a requirement for long period, since many Sri Lankans use this writing method to write product reviews, social media posts and comments etc. This has been tried since couple of years by many research students but the main challenge was to find a proper data set to evaluate deep learning models for this Natural Language Processing (NLP) task. Hence, traditional statistic, rule-based models has been proposed with less data. This research addresses the challenge of preparing a data set to evaluate a deep learning approach for this machine translation activity and also to evaluate a seq2seq Neural Machine Translation (NMT) model. The proposed seq2seq model is purely based on the attention mechanism, as it has been used to improve NMT by selectively focusing on parts of the source sentence during translation. The proposed approach can achieve 24.13 BLEU score on Singlish-English by seeing ~0.15 M parallel sentence pairs with ~50 K word vocabulary.

Keywords:
Singlish, NMT, Language processing, seq2seq, Attention model, word embedding.

Dedication

I dedicate my dissertation to my family and many friends. A special feeling of gratitude is extended to my loving parents whose thoughts of encouragement supported me in reaching this milestone. I will always appreciate the support of my friends for all the things they have done on behalf of me and their valuable thoughts. Last but not least, I dedicate this work with heartfelt gratitude to my supervisors and the staff of University of Moratuwa for their help, specially to the academic staff for providing me guidance throughout this research.

Above all, I am ever indebted, like every other Sri Lankan, to all those citizens who supported Free Education with their taxes and to all those students who fought with their life to protect Free Education in Sri Lanka.

Acknowledgement

First and foremost, I acknowledge my supervisors Dr. Sagara Sumathipala and Dr. Subha Fernando for their insight for making my research successful. Without the support from supervisors, it would undoubtedly be a real challenge for me to make this research a success.

I would like to thank University of Moratuwa for giving an opportunity to carry out this research and its continuing support during the research. I would like to extend my gratitude for all the academic and non-academic staff of the University of Moratuwa and my colleagues for their generous support, comments and encouragement throughout the project. I am ever grateful to all the expert and novice meditators who were involved in this research project.

Table of Contents

List of Figures

List of Tables

List of Abbreviation

NLP – Natural Language Processing

NMT – Neural Machine Translation

MT – Machine Translation

RBMT – Rule Based Machine Translation

BERT – Bidirectional Encoder Representations from Transformers

BLEU – Bilingual Evaluation Understudy

RNN – Recurrent Neural Network

LSTM – Long Short Term Memory

SGD – Stochastic Gradient Decent

TF-IDF – Term Frequency-Inverse Document Frequency