Using Back-Translation to improve domain-specific English-Sinhala Neural Machine Translation

Koshiya Epaliyana

208038N

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree Master of Science in Computer Science and Engineering

Department of Computer Science & Engineering

University of Moratuwa

Sri Lanka

June 2021

DECLARATION

I, Koshiya Epaliyana, declare that this is my own work and this dissertation does not incorporate without acknowledgment any material previously submitted for a Degree or Diploma in any other University or institute of higher learning, and to the best of my knowledge and belief, it does not contain any material previously published or written by another person except where the acknowledgment is made in the text.

Also, I hereby grant to the University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:

The above candidate has carried out research for the Masters thesis/Dissertation under my supervision.

Name of Supervisor: Dr. Surangika Ranathunga

Signature of the Supervisor: Date:

Name of Supervisor: Prof. Sanath Jayasena

Signature of the Supervisor: Date:

ABSTRACT

Machine Translation (MT) is the automatic conversion of text in one language to other languages. Neural Machine Translation (NMT) is the state-of-the-art MT technique w builds an end-to-end neural model that generates an output sentence in a target language given a sentence in the source language as the input.

NMT requires abundant parallel data to achieve good results. For low-resource settings such as Sinhala-English where parallel data is scarce, NMT tends to give sub-optimal results. This is severe when the translation is domain-specific. One solution for the data scarcity problem is data augmentation. To augment the parallel data for low-resource language pairs, commonly available large monolingual corpora can be used. A popular data augmentation technique is Back-Translation (BT). Over the years, there have been many techniques to improve vanilla BT. Prominent ones are Iterative BT, Filtering, Data Selection, and Tagged BT. Since these techniques have been rarely used on an inordinately low-resource language pair like Sinhala - English, we employ these techniques on this language pair for domain-specific translations in pursuance of improving the performance of Back-Translation. In particular, we move forward from previous research and show that by combining these different techniques, an even better result can be obtained. In addition to the aforementioned approaches, we also conducted an empirical evaluation of sentence embedding techniques (LASER, LaBSE, and FastText+VecMap) for the Sinhala-English language pair.

Our best model provided a +3.24 BLEU score gain over the Baseline NMT model and a +2.17 BLEU score gain over the vanilla BT model for Sinhala → English translation. Furthermore, a +1.26 BLEU score gain over the Baseline NMT model and a +2.93 BLEU score gain over the vanilla BT model were observed for the best model for English → Sinhala translation.

Keywords: Neural Machine Translation, Back-Translation, Data selection, Iterative Back-Translation, Iterative filtering , Low-resource languages, Sinhala

ACKNOWLEDGEMENTS

To start with, I would like to convey my sincere gratitude to my supervisors Dr. Surangika Ranathunga and Professor Sanath Jayasena for the tremendous support and guidance they provided me with, throughout the entire period of the research. I'm grateful for your insights, advice, and encouragement. Without your guidance and support, I could not have achieved this milestone. Your thorough knowledge of Machine Learning, Natural Language Processing, and Deep learning continuously helped me to push myself to learn more, dig deep into study material, and try out new techniques/approaches.

I wish to thank Prof. Gihan Dias for his valuable insights and guidance from the early stage of this research. I would also like to thank both the academic and non-academic staff of the Department of Computer Science and Engineering, for providing me with the resources necessary to conduct my research. This research was supported by the University of Moratuwa AHEAD project Research Grant.

Finally, I would like to give my thanks to my friends and family for all their love and support.

Thank you!

LIST OF ABBREVIATIONS

MT	Machine Translation
NLP	Natural Language Processing
NMT	Neural Machine Translation
SMT	Statistical Machine Translation
BT	Back-Translation
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
LSTM	Long Short Term Memory
RBMT	Rule Based Machine Translation
FDA	Feature Decay Algorithm
INR	Infrequent n-gram Recovery
RCTM	Recurrent Continuous Translation Model
RNNEncdec	RNN Encoder-Decoder

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS