DATA AUGMENTATION TO INDUCE HIGH QUALITY PARALLEL DATA FOR LOW-RESOURCE NEURAL MACHINE TRANSLATION

W.A.S.A Fernando

208035D

Doctor of Philosophy

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

September 2025

DATA AUGMENTATION TO INDUCE HIGH QUALITY PARALLEL DATA FOR LOW-RESOURCE NEURAL MACHINE TRANSLATION

W.A.S.A Fernando

208035D

Dissertation submitted in partial fulfillment of the requirements for the degree
Doctor of Philosophy

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

September 2025

DECLARATION

I declare that this is my own work and this Dissertation does not incorporate without acknowledgement 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 acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 03.09.2025

The supervisors should certify the Dissertation with the following declaration.

The above candidate has carried out research for the Doctor of Philosophy Dissertation under our supervision. We confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr. Nisansa de Silva

Signature of the Supervisor: Digitally signed by Nisansa de Silva Date: 2025.09.03 16:15:47 +05'30' Date: 03.09.2025

Name of Supervisor: Dr. Surangika Ranathunga

Signature of the Supervisor: Surangika Ranathunga Digitally signed by Surangika Ranathunga Date: 2025.09.03 22:07:46 +12'00' Date: 03.09.2025

DEDICATION

To *The God*

For the blessings throughout my life and for aligning
this opportunity for me.

To My *Thaththa*, Ranjith and *Amma*, Indrani,

For your selfless love and endless sacrifices for shaping me into the
person I am today.

To My *Husband*, Kumudu and
our *Children*, Kaviru, Adeesha and Asheth,

For your endless love, encouragement and strength you have
always given me.

ACKNOWLEDGEMENT

I am grateful for my supervisors, Dr Nisansa de Silva and Dr Surangika Ranathunga, for the unreserved guidance and support provided during the course of my PhD journey. I highly appreciate Dr. Surangika Ranathunga for inspiring me to take this path and for continuing to mentor me with dedication, even after her transition to another University. I extend the same appreciation to Dr. Nisansa de Silva, for his willingness be my supervisor and for the invaluable mentorship since then. I consider myself truly fortunate to have had the opportunity to work under their supervision.

I would like to sincerely thank Prof. Gihan Dias for giving me the opportunity to join the National Languages Processing Centre (NLPC) at the University of Moratuwa as a Research Engineer. This opportunity laid the foundation for my research. I am also grateful for his encouragement and unwavering support throughout this academic journey.

I wish to thank Dr Uthayasanker Thayasivam, current Head of the Department of Computer Science and Engineering for his unresearved support extended during this time, and for his continuous motivation towards completing the PhD.

I am thankful to Prof. Sanath Jayasena and Dr. Kutila Gunasekara, CSE Research Coordinator, for their involvement and helpful contributions during my academic journey.

I am grateful to my progress panel, Prof. Asoka Karunananda, Dr. Charith Chitraranjan, and Dr. Lochandaka Ranathunga, for their valuable feedback and insightful suggestions that helped enrich and strengthen my research. I am thankful to all the lecturers at the Department of Computer Science and Engineering for extending their support throughout this time.

Further, I would like to acknowledge the Accelerating Higher Education Expansion and Development (AHEAD) Operation of the Ministry of Education, funded by the World Bank for facilitating the initial part of the research. Secondly, I wish to acknowledge the Senate Research Committee (SRC) grant of the University of Moratuwa, Sri Lanka, for funding the second part of the research. I wish to acknowledge the LK domain registry for funding me with the Prof. V. K. Samaranayake top-up grant during the third phase of this research. The final phase of this research was funded by the Google Award for Inclusion Research (AIR) 2022 received by Dr Surangika Ranathunga and Dr Nisansa de Silva. I would like to thank and acknowledge the National Languages Processing (NLP) Centre, at the University of Moratuwa for providing the GPUs to execute the experiments related to the research.

ABSTRACT

Supervised Neural Machine Translation (NMT) models rely on parallel data to produce reliable translations. A parallel dataset refers to a collection of texts in two or more languages in which each sentence in one language is aligned with its corresponding translation counterpart in the other language. NMT models have produced state-of-the-art results for High-Resource Languages. HRLs refer to languages that have linguistic resources and tool support. In Low-Resource settings, which means for languages with limited or no linguistic resources and/or tools, NMT performance is suboptimal due to two challenges: the parallel data scarcity problem and the presence of noise in the available parallel datasets. Data augmentation (DA) is a viable approach to address these problems, as it aims to induce *high-quality* parallel sentences efficiently using automatic means. In our research, we begin by analysing the limitations of the existing DA methods and propose strategies to overcome those limitations, aimed at improving the NMT performance. To generalise our findings, we conduct this study on three language pairs: English-Sinhala, English-Tamil and Sinhala-Tamil. They belong to three distinct language families. Further, Sinhala and Tamil are known to be morphologically rich languages, making NMT further challenging.

First, we follow the word or phrase replacement-based augmentation strategy, where we induce *synthetic* parallel sentences by augmenting rare words and by using words from a bilingual dictionary. Our technique improves existing techniques by using both syntactic and semantic constraints to generate high-quality parallel sentences. This method improves translation quality for sequences containing out-of-vocabulary terms and yields better overall NMT scores than existing techniques. Secondly, we conduct an empirical study with three multilingual pre-trained language models (multiPLMs) and demonstrate that both the pre-training strategy and the nature of the pre-training data significantly affect the quality of mined parallel sentences. Thirdly, we enhance the cross-lingual sentence representations of existing encoder-based multiPLMs, in order to overcome their suboptimal performance in sentence retrieval tasks. We introduce *Linguistic Entity Masking* to enhance the cross-lingual representations of such multiPLMs and empirically prove that the improved representations lead to performance gains for cross-lingual tasks. Finally, we explore the Parallel Data Curation (PDC) task. In line with existing work, we identify that the scoring and ranking with different multiPLMs results in a disparity, which is caused by the multiPLMs' bias towards certain types of noisy parallel sentences. We show that multiPLM-based PDC, together with a heuristic combination, is capable of minimising this disparity while producing optimal NMT scores. Overall, we show that improving DA techniques leads to generating *high-quality* parallel data, which in turn leads to elevating the state-of-the-art benchmark NMT results further.

Keywords: Data Augmentation, Neural Machine Translation, Low Resource Languages, Bitext Mining, Parallel Data Curation

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

Abbreviation	Description
AI	Artificial Intelligence
BPE	Byte-pair-Encoding
CNN	Convolutional Neural Network
DA	Data Augmentation
DAN	Deep Averaging Networks
DNN	Deep Neural Network
GRU	Gated Recurrent-Unit
HRL	High-Resource Languages
LLM	Large Language Models
LM	Language Model
LRL	Low Resource Languages
LRNMT	Low Resource Neural Machine Translation
LSTM	Long Short-Term Memory
MNMT	Multilingual Neural Machine Translation
MT	Machine Translation
multiPLM	Multilingual Pre-trained Language Model
NER	Named Entity Recognition
NET	Named Entity Translation
NLP	Natural Language Processing
NMT	Neural Machine Translation
OOV	Out-of-Vocabulary
PBSMT	Phrase-Based Statistical Machine Translation
POS	Part of Speech
RNN	Recurrent Neural Networks

Abbreviation	Description
RBMT	Rule-based Machine Translation
SMT	Statistical Machine Translation
UNMT	Unsupervised Neural Machine Translation