

LB/TH/38/2025
TH5965

**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:

 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 unreserved 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

Declaration of the Candidate & Supervisor	i
Dedication	ii
Acknowledgement	iii
Abstract	iv
Table of Contents	v
List of Figures	x
List of Tables	xii
Abbreviations	xvi
1 Introduction	1
1.1 Background and Motivation	1
1.2 Research Objectives	2
1.3 Contributions	5
1.4 Structure of the Thesis	7
1.5 Publications	8
1.6 Other Publications	8
1.7 Definitions	9
2 Background	11
2.1 Machine Translation	11
2.2 Machine Translation Techniques	11
2.2.1 Rule-based Machine Translation	11
2.2.2 Statistical Machine Translation	12
2.2.3 Neural Machine Translation (NMT)	13
2.3 Prominent Techniques in NMT	17
2.4 Low-Resource Neural Machine Translation	18
2.5 Data Augmentation	19
2.5.1 Word or Phrase Replacement-based augmentation	19
2.5.2 Bitext Mining	20

2.5.3	Parallel Data Curation	20
2.5.4	Back-Translation	20
2.6	Sinhala-Tamil-English Related NMT	21
2.6.1	Selection of Low-Resource Language Pairs	21
2.6.2	Progression NMT Research among Sinhala-Tamil-English Languages	22
2.6.3	Dataset Availability	23
2.7	Chapter Summary	23
3	Generating Synthetic Parallel Sentences	24
3.1	Introduction	24
3.2	Related Work	25
3.3	Methodology	26
3.3.1	Rare Word Augmentation	26
3.3.2	Dictionary Augmentation	28
3.3.3	Combined Solution	29
3.4	Experiments	29
3.4.1	Dataset	29
3.4.2	NMT Experiment Setup	30
3.4.3	Baseline Models	31
3.4.4	Augmentation of Rare Words	32
3.4.5	Augmentation of Dictionary	32
3.4.6	Combined Experiments	33
3.5	Results Analysis	33
3.5.1	Rare Word Augmentation	33
3.5.2	Dictionary Augmentation	34
3.5.3	Qualitative Analysis	35
3.5.4	NMT Results on Transformer Architecture	35
3.6	Discussion	36
3.7	Chapter Summary	37

4	Empirical Study: multiPLMs for Bitext Mining	39
4.1	Introduction	39
4.2	Related Work	40
4.2.1	Document Alignment	40
4.2.2	Sentence Alignment	41
4.2.3	Pre-trained Multilingual Language Models (multiPLMs)	42
4.2.4	Evaluating Document Alignment and Sentence Alignment Tasks	42
4.3	Methodology	42
4.3.1	Dataset	42
4.3.2	Justification for Selecting MultiPLMs	45
4.3.3	Document Alignment	46
4.3.4	Sentence Alignment	50
4.4	Experiments and Results	52
4.4.1	Document Alignment	52
4.4.2	Sentence Alignment	56
4.4.3	Extrinsic Evaluation with NMT	58
4.5	Discussion	60
4.6	Chapter Summary	62
5	Linguistic Entity Masking (LEM)	63
5.1	Introduction	63
5.2	Motivation	64
5.3	Related Work	65
5.3.1	MLM and TLM Objectives	65
5.3.2	Different Masking Strategies	65
5.4	Methodology	66
5.5	Theoretical Framework for Linguistic Entity Masking (LEM)	67
5.6	Experiments	69
5.6.1	Impact of the type of monolingual data in LEM_{mono}	70
5.6.2	Evaluation of Different Masking Strategies	70
5.6.3	Evaluation of LEM Strategy and Ablation Study	70
5.6.4	Evaluation Tasks	71

5.7	Experiment Setup	72
5.7.1	Data Selection	72
5.7.2	MultiPLM Selection	73
5.7.3	Baselines	74
5.7.4	Implementation and Hyper-parameters	74
5.8	Experimental Results	76
5.8.1	Impact of the type of monolingual data in LEM_{mono}	76
5.8.2	Evaluation of Different Masking Strategies	76
5.8.3	Evaluation of LEM Strategy and Ablation Study	77
5.8.4	Parallel Data Curation	81
5.9	Ablation Studies	82
5.9.1	The Number of Tokens for Masking in LEM Strategy	82
5.9.2	Effect of noise in LEM Strategy	83
5.10	Discussion	84
5.11	Chapter Summary	85
6	Debiasing the Disparity in NMT systems	86
6.1	Introduction	86
6.2	Motivation	86
6.3	Related Work	89
6.3.1	MultiPLMs for PDC	89
6.3.2	Identifying Noise in Web-mined Corpora	90
6.3.3	Heuristic-based PDC	90
6.4	Methodology	91
6.4.1	Improved Taxonomy for Noise	91
6.4.2	Selection of Heuristics	91
6.4.3	Human Evaluation	91
6.5	Experiments	92
6.5.1	Dataset	93
6.5.2	Selection of multiPLMs	93
6.5.3	Heuristic-based PDC Experiments	93
6.5.4	NMT Experiments	94

6.6	Experimental Results	94
6.6.1	Impact of Heuristics on NMT Results	96
6.6.2	Summary of Heuristic-based PDC	98
7	Discussion	99
7.1	Research Objectives	99
7.2	Future Work	101
7.2.1	Inducing Synthetic Sentences	101
7.2.2	Effectiveness of multiPLMs on Document Alignment and Sentence Alignment tasks	102
7.2.3	Improving Cross-Lingual Representations	102
7.2.4	Parallel Data Curation	102
7.3	Chapter Summary	103
8	Conclusion	104
	References	105
	Appendix A Monolingual Data for MLM Step	130
	Appendix B PDC: Extrinsic Evaluation Results	131
	Appendix C Debiasing Disparity with Heuristics	132
	C.1 Improved Taxonomy for Noise Categorization	132
	C.2 Human Evaluation	132

LIST OF FIGURES

Figure	Description	Page
Figure 1.1	Research objectives vs contributions and publication mapping	3
Figure 2.1	Classification of MT Techniques	12
Figure 2.2	Translation example with RNN. Adapted from (Sutskever et al., 2014)	14
Figure 2.3	Transformer Architecture. Adapted from Zhang and Zong (2020)	15
Figure 2.4	Classification of Data Augmentation Techniques in NMT	19
Figure 3.1	Data Augmentation Process.	26
Figure 3.2	Shows the limitation with the word alignment (GIZA++) model and the limitation with the morphological analyser.	38
Figure 4.1	Process for calculating the semantic distance between source language document d_A and target language document d_B . Here $w_{A,B}$ refers to the weight considering bilingual lexicons between sentence s_A and s_B . The semantic distance scored from this process would be used by the Document matching algorithm (Section 4.3.3.1) to finally produce the aligned document pairs.	48
Figure 4.2	Given the source and target language sentences, the diagram outlines the sentence alignment algorithm considering the forward criterion. In the backwards criterion, for each s_B in d_B , the aligned sentences are picked up from the source side.	51
Figure 5.1	Self-attention weights among the words for an English and its corresponding Sinhala sentence. The darker the colour is, the stronger the relationship (ie. self-attention weight) between the two words.	64
Figure 5.2	A comparison of existing masking strategies is presented using an example from the English-Sinhala language pair. Sub-word masking, Whole Word masking, span masking, and LEM_{mono} exclusively utilize monolingual sentences during masking. In contrast, TLM and LEM_{para} apply masking on concatenated parallel sentences. Notably, in both LEM_{mono} and LEM_{para} , only a single token from the linguistic entity is masked.	67
Figure 5.3	The LEM continual pre-training process. An existing <i>multiPLM</i> , is first continually pre-trained (LEM_{mono}) with <i>dependent monolingual data</i> . In the second continual pre-training step (LEM_{para}), the LEM strategy is applied on the <i>concatenated parallel data</i> .	68

Figure 5.4	Sentence alignment Recall scores for using independent monolingual data (MADLAD-400) versus dependent monolingual data obtained from the parallel corpus (SiTa-Trilingual parallel Corpus). Here the Forward (FW), Backward (BW) and Intersection (IN) approach refers to the criterion followed to identify the translation sentences as per the work of Artetxe and Schwenk (2019a).	76
Figure 6.1	Baseline NMT scores in ChrF++ when trained with the top-ranked sentence pairs from CCMatrix and CCAIined, using embeddings obtained from LASER3, XLM-R, and LaBSE.	87
Figure 6.2	Percentage of <i>dedup+ngram</i> experiments exceeding the best result of <i>dedup</i> for each <i>multiPLM</i>	96
Figure 6.3	Percentage of <i>dedup+ngram</i> experiments exceeding the best result of <i>dedup</i> with respect to the Language-pair.	97
Figure C.1	Shows the annotation guideline document in terms of a flow chart. This shows the priority of the noise category to be selected prior to declaring the annotation class.	133

LIST OF TABLES

Table	Description	Page
Table 3.1	Parallel Corpus Statistics of Training and Validation sets	29
Table 3.2	Test set Statistics	30
Table 3.3	Statistics corresponding to the Monolingual Corpus to train the LMs.	30
Table 3.4	Rare Word Augmentation Results considering different syntactic and semantic constraints	34
Table 3.5	Dictionary Word Augmentation Results considering different syntactic and semantic constraints. Here, TS1, TS2 and TS3 correspond to the three evaluation sets.	35
Table 3.6	Improvement in the En translation with respective to each augmented dataset. The input S_i sentence contains <i>parisilanaya</i> , the OOV term. Using more syntactic and semantic constraints improves the fluency and completeness of the translated sentence.	36
Table 3.7	NMT Results on transformer architecture, trained with the best performing syntactic and semantic combination from the rare word and dictionary term augmentation experiments.	37
Table 4.1	Statistics of document alignment evaluation dataset	43
Table 4.2	Statistics of the sentence alignment evaluation dataset	44
Table 4.3	Statistics of the Bilingual Lexicons	44
Table 4.4	Overview of the Bilingual Lexicons	45
Table 4.5	Overview of the Improved Dictionary	50
Table 4.6	Document Alignment results in terms of Precision(P), Recall (R) and F1 for English-Sinhala language pair.	53
Table 4.7	Document Alignment results in terms of Precision(P), Recall (R) and F1 for English-Tamil language pair.	54
Table 4.8	Document Alignment results in terms of Precision (P), Recall (R) and F1 for Sinhala-Tamil language pair.	55
Table 4.9	Sentence Alignment Results in terms of Recall (R). Here, B refers to the score obtained using Artetxe and Schwenk (2019a)’s method and $B+D$ refers to the scores obtained using Rajitha et al. (2020) bilingual lexicon improvement.	57
Table 4.10	BLEU Scores for NMT systems trained with parallel data obtained from Sentence Alignment step with Forward (F), Backward (B) and Intersection (I) criterion	59

Table 4.11	Error Analysis in the sentence alignment task. Here, the alignment[corr] refers to the alignment in the gold-standard evaluation set and alignment[incorr] refers to the alignment produced in the experiments.	61
Table 5.1	Existing masking strategies. The <i>Masked Token Type</i> indicates the type of words considered for masking. We include our masking strategy (LEM) for comparison purposes.	66
Table 5.2	English (En), Sinhala (Si), and Tamil (Ta) examples of the returned sub-words after the tokenization step are presented. In English, nouns are typically inflected based solely on number. In contrast, Sinhala and Tamil nouns undergo inflection not only based on number but also on case category and gender.	69
Table 5.3	Hyper-parameters used during continual pertaining with LEM strategy	74
Table 5.4	Training parameters for NMT experiments.	75
Table 5.5	Sentence alignment Recall scores for the different masking strategies.	77
Table 5.6	Ablation experiments and sentence alignment scores for English-Sinhala language pair considering linguistic entity masking.	78
Table 5.7	Ablation experiments and sentence alignment scores for English-Tamil language pair considering linguistic entity masking.	79
Table 5.8	Ablation experiments and sentence alignment scores for Sinhala-Tamil language pair considering linguistic entity masking.	80
Table 5.9	Results for sentence alignment task in terms of recall points. For comparison purposes, the FW, BW and IN gains are averaged and reported in the <i>Overall Average Gain column</i> .	81
Table 5.10	ChrF++ scores for the parallel data curation task. The scores have been reported on the Flores+ devtest. The values in brackets indicate the gains of XLM-R _{LEM} compared to the XLM-R and the XLM-R _{MLM+TLM} respectively.	82
Table 5.11	The Recall scores from the ablation study by changing the number of tokens masked in the linguistic entity. The results are for the Sinhala-Tamil language pair and the sentence alignment downstream task.	83
Table 5.12	Sentence alignment Recall results obtained using LEM-enhanced models on both high-quality and noisy web-crawled datasets.	84
Table 5.13	Examples of incorrect identification and labeling of NEs. We identify two error categories: false positives and false negatives, where the NER model underperforms.	84
Table 5.14	Examples of incorrect identification and labelling of POS Tags. We identify mainly two error categories: false positives and false negatives, where the Pos Tagger underperforms.	85
Table 6.1	Example parallel sentences from the En-Si, En-Ta and Si-Ta, identified during human evaluation.	88

Table 6.2	Human evaluation results for CCMatrix and CCAIined for En-Si, En-Ta and Si-Ta. Results are reported for LASER3, XLM-R, and LaBSE before and after applying heuristics. We report the average score among the scores obtained from the individual annotators. (C) - overall correct percentage considering CC (perfect translation), CN (near perfect) and CB (boilerplate). (E) - overall error percentage considering CCN (correct with overlaps), CS (correct but short sentence), X (wrong translation), UN (untranslated), WL (wrong language), NL (not a language).	88
Table 6.3	Noise types in parallel corpora, as identified by Khayrallah and Koehn (2018) (A) , Bane et al. (2022) (B) , Herold et al. (2022) (C) , Kreutzer et al. (2022b) (D) and Ranathunga et al. (2024a) (E) .	89
Table 6.4	A comparison of the improved taxonomy against Ranathunga et al. (2024a)'s. (only showing the changes)	91
Table 6.5	Mapping between the noise category vs the noise mitigating heuristic.	92
Table 6.6	Corpus statistics.	93
Table 6.7	Training parameters for NMT experiments.	95
Table 6.8	NMT results obtained after applying heuristics in isolation and in combination in the ablation study. The values in bold indicate the highest NMT score obtained for a given heuristic class or from the heuristic combination. The values underlined are the highest among the individual heuristics. Highlighted in green are the overall best values. Here DD+PN is <i>Deduplication+punctNums</i> , SL is <i>sLength</i> and LT is <i>LIDThresh</i> . Here NA would be when the particular experiment is not applicable for that language pair or the dataset.	95
Table A.1	Bitext mining recall scores for using pure monolingual data versus source and target sides from a parallel corpus (as monolingual data) for MLM experiments.	130
Table B.1	NMT scores on the Flores+ devtest using top 50,000 parallel sentences from the ranked NLLB and CCAIined corpus.	131
Table C.1	Example parallel sentences which will be separately identified under the new noise category <i>CCN</i>	132
Table C.2	Ranathunga et al. (2024a)'s error taxonomy with the CCN category that has been newly added by us	132
Table C.3	Annotator details with the years of experience and their qualifications.	133
Table C.4	Shows the final corpus sizes after applying heuristics, along with the reduction percentage. Here DD+PN is <i>Deduplication+punctNums</i> , SL is <i>sLength</i> and LT is <i>LIDThresh</i> . NA corresponds to the experiments that are not applicable for the language pair. Red(%) refers to the percentage reduction of the dataset size due to applying the heuristics.	134

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