

# 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 which 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  $\rightarrow$  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  $\rightarrow$  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

Figure 2.1	Back-Translation process	7
Figure 2.2	Iterative Back-Translation process	9
Figure 3.1	Research process	25
Figure 3.2	En $\rightarrow$ Si Filtered BT process	28
Figure 4.1	Pre-processing script	34
Figure 4.2	Training script	35
Figure 5.1	<b>Si</b> $\rightarrow$ <b>En</b> vanilla BT model and the best Filtered BT models from each embedding technique	44
Figure 5.2	<b>Si</b> $\rightarrow$ <b>En</b> Filtered BT for different threshold values	45
Figure 5.3	<b>En</b> $\rightarrow$ <b>Si</b> vanilla BT model and the best Filtered BT models from each embedding technique	47
Figure 5.4	<b>En</b> $\rightarrow$ <b>Si</b> Filtered BT for different thresholds	48
Figure 5.5	Sentence pairs picked by different embedding techniques	49
Figure 5.6	Sentence pairs filtered-out by different embedding techniques	49
Figure 5.7	<b>Si</b> $\rightarrow$ <b>En</b> Iterative BT and Iterative Filtered BT	51
Figure 5.8	<b>Si</b> $\rightarrow$ <b>En</b> FDA and INR combined with Iterative BT and Iterative Filtered BT	52
Figure 5.9	<b>En</b> $\rightarrow$ <b>Si</b> Iterative BT and Iterative Filtered BT	53
Figure 5.10	<b>En</b> $\rightarrow$ <b>Si</b> FDA and INR combined with Iterative BT and Iterative Filtered BT	54
Figure 5.11	Monolingual target sentences selected by FDA and INR algorithms	55
Figure 5.12	Monolingual target sentences rejected by FDA and INR algorithms	55
Figure 5.13	Different authentic to synthetic parallel data ratio (News data)	59

## LIST OF TABLES

Table 3.1	<b>Tagged synthetic sentence</b>	31
Table 4.1	<b>Parallel Data for Si-En</b>	37
Table 4.2	<b>Monolingual Data</b>	37
Table 4.3	<b>Monolingual Data for different ratios in the News domain</b>	38
Table 5.1	<b>Si→En</b> Vanilla BT	40
Table 5.2	<b>En→Si</b> Vanilla BT	41
Table 5.3	Iterative BT with Data selection for <b>Si → En</b>	42
Table 5.4	Iterative BT with Data selection for <b>En → Si</b>	42
Table 5.5	<b>Si→En</b> Filtered BT with different threshold values and embedding techniques	43
Table 5.6	<b>En→Si</b> Filtered BT with different threshold values and embedding techniques	46
Table 5.7	Iterative Filtered BT (different embedding techniques) with Data selection for <b>Si → En</b>	50
Table 5.8	Iterative Filtered BT (different embedding techniques) with Data selection for <b>En → Si</b>	52
Table 5.9	Tagged BT and Iterative Tagged BT for <b>Si → En</b>	56
Table 5.10	Tagged BT with filtering and Iterative Tagged BT with filtering (LASER) for <b>Si → En</b>	56
Table 5.11	Tagged BT and Iterative Tagged BT for <b>En → Si</b>	57
Table 5.12	Tagged BT with filtering and Iterative Tagged BT with filtering (LASER) for <b>En → Si</b>	57
Table 5.13	Performance with different authentic to synthetic data ratios	58
Table 5.14	<b>Filtering with different thresholds with LASER as the sentence embedding technique.</b>	60
Table 5.15	<b>Iterative Filtered BT with Data selection (All the models are Ensemble models)</b>	60
Table 5.16	<b>Best models for Si → En</b>	64





# TABLE OF CONTENTS

Declaration of the Candidate & Supervisor	i
Abstract	ii
Acknowledgement	iii
List of Abbreviations	iv
List of Figures	v
List of Tables	vi
Table of Contents	viii
1 Introduction	1
1.1 Background	1
1.2 Research Problem	2
1.3 Research Objectives	3
1.4 Contributions	4
1.5 Publications	5
1.6 Organization	5
2 Literature Survey	6
2.1 Basic Back-Translation	6
2.2 Iterative Back-Translation	8
2.3 Data Selection	10
2.3.1 Transductive Data Selection algorithms	11
2.3.2 Selecting sentences consisting of difficult to predict words	13
2.3.3 Selecting target domain data	13
2.4 Filtering	14
2.4.1 Sentence-level similarity metrics using surface information of sentences	14
2.4.2 Sentence-level similarity metrics using distributed represen- tations of sentences	14
2.4.3 Comprehensive analysis of different filtering techniques	15
2.4.4 Sentence Embedding techniques	17

2.5	Tagged BT	18
2.6	Other approaches	19
2.6.1	Using both target-side and source-side monolingual data	19
2.6.2	Noised Back-Translation	20
2.6.3	Sampling	21
2.6.4	Using a pivot language	21
2.6.5	Training the model on synthetic data and fine-tuning on authentic data	22
2.6.6	The impact of the size of the monolingual corpus on Back-Translation	22
2.7	Summary	23
3	Methodology	24
3.1	Vanilla Back-Translation	25
3.2	Iterative Back-Translation	26
3.3	Filtered BT	26
3.3.1	Iterative Filtered Back-Translation	29
3.4	Data selection	29
3.5	Iterative Filtered BT with Data selection	30
3.6	Tagged Back-Translation	31
3.6.1	Iterative Tagged Back-Translation	31
3.6.2	Tagged BT with Filtering	32
3.6.3	Iterative Tagged BT with Filtering	32
4	Experiments	34
4.1	Setup	34
4.2	Baseline NMT model	34
4.3	Data	36
4.4	Experimental details	37
5	Results and Discussion	40
5.1	In-domain data	40
5.1.1	Vanilla BT	40
5.1.2	Iterative BT	41

5.1.3	Filtered Back-Translation	42
5.1.4	Iterative Filtered BT with Data Selection	50
5.1.5	Tagged Back-Translation	55
5.2	Out-of-domain data	58
5.2.1	Selecting the best authentic to synthetic data ratio	58
5.2.2	Filtered BT and Iterative Filtered BT with Data Selection	59
5.3	Discussion	61
5.3.1	In-domain data	61
5.3.2	Out-of-domain data	63
5.4	Best 5 models obtained for each translation direction	64
6	Conclusion and Future work	66
	References	68