# Bilingual Lexical Induction for English-Sinhala

Anushika Liyanage

208033U

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

Department of Computer Science & Engineering

University of Moratuwa
Sri Lanka

October 2022

# DECLARATION

I, Anushika Liyanage, 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.

Also, I hereby grant to 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: UOM Verified Signature Date: 07/10/2022

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

Name of Supervisor: Dr. Surangika Ranathunga

Signature of the Supervisor: Surangika Ranathunga UOM Verified Signature Digitally signed by Surangika Ranathunga Date: 2023.02.04 12:40:33 +05'30' Date: 01/02/2023

Name of Supervisor: Dr. Sanath Jayasena

Signature of the Supervisor: UOM Verified Signature Date: 01/02/2023

# ABSTRACT

**Bilingual Lexicons** are important resources appertaining to Natural Language Processing (NLP) applications such as Neural Machine Translation and Named Entity Recognition (NER). However, Low Resource Languages (LRLs) equivalent to Sinhala lack such resources. Manually producing millions of word translations between languages is exhaustive and almost impossible. An increasingly popular approach to automatically create such resources is Bilingual Lexical Induction (BLI).

We created the first-ever BLI model for English and Sinhala language pair using the existing popular model VecMap. Currently, no prior work has conducted a sufficient evaluation with respect to the factors, nature of the dataset, type of embedding model used, or the type of evaluation dictionary used on BLI and how these factors affect the results of BLI. We fill the gap by executing an extensive set of experiments with regard to the aforementioned factors on BLI for Sinhala and English in this thesis.

Furthermore, we enhance the pre-trained embeddings to cater to the application by applying sophisticated post-processing approaches. Linear transformation and effective dimensionality reduction are applied to the pre-trained embeddings before obtaining cross-lingual word embeddings between Sinhala and English by applying VecMap. Furthermore, we have introduced dimensionality reduction to the VecMap algorithm where the algorithm starts the first iteration from a low dimension to initialize a better solution. Subsequently, the dimensionality of the embeddings is increased in each iteration until embeddings reach the original dimension in the final iteration. We were able to improve the results considerably by learning a better initial solution and hence an improved final solution. Finally, we combined the post-processing step with the modified VecMap model to obtain even better mapping for Sinhala-English language pair which in turn is applicable in task-specific downstream systems to improve the results of the entire system.

**Keywords**: Sinhala; embedding spaces; embedding models; bilingual lexicon induction

## ACKNOWLEDGEMENTS

I would like to recognize the invaluable guidance of my supervisors, Dr.Surangika Ranathunga and Prof.Sanath Jayasena. I have greatly benefited from their insights, vast knowledge, skillful supervision, and suggestions made throughout this research. I am very thankful for the constant support, patience, understanding, and motivation extended to me when times were difficult.

I would like to thank Prof. Gihan Dias, for the critical advice given throughout this research. I would like to extend my gratitude to the academic and non-academic staff of the Department of Computer Science and Engineering for providing the required resources to carry out this research. This research was funded by the University of Moratuwa AHEAD project research grant.

I thank my partner Milinda Rajapakshe for listening to my constant rant, unconditional love, support, and for the sacrifices you made in order for me to pursue a master's degree. I am forever grateful to my friends and family for the love, support, and motivation in this entire thesis process every day.

**Thank you!**

# LIST OF ABBREVIATIONS

| | |
|---|---|
| NLP | Natural Language Processing |
| NMT | Neural Machine Translation |
| NER | Named Entity Recognition |
| BLI | Bilingual Lexical Induction |
| PCA | Principal Component Analysis |
| PPA | Post Processing Algorithm |
| RNN | Recurrent Neural Networks |
| SA | Sentiment Analysis |
| LRL | Low Resource Languages |
| HRL | High Resource Languages |
| SVD | Singular Value Decomposition |
| NN | Nearest Neighbor |

## LIST OF FIGURES

# LIST OF TABLES

# TABLE OF CONTENTS