

# 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: 07/10/20  
22

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: *UOM Verified Signature*

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

Figure 2.1	Illustrative example of vector representation of words in a two-dimensional space	7
Figure 2.2	Simple word vector representation matrix where each row of the matrix represents a word vector and columns represent the dimensions	7
Figure 2.3	Architecture of the word2vec models: CBOW and Skip-Gram. Image source [1]	8
Figure 2.4	Geometric Similarity between English and Spanish. Image source [2]	10
Figure 2.5	VecMap Model. Image source [3]	17
Figure 3.1	Research Process	23
Figure 5.1	Translation Difference Examples	41

## LIST OF TABLES

Table 4.1	Dataset	36
Table 5.1	BLI Results For Pre-trained fastText Embeddings	40
Table 5.2	BLI Results For Combined News Data	41
Table 5.3	BLI Results For Each News Source Separatelyfit to pg width	42
Table 5.4	BLI Results For Few News Sources Combined	45
Table 5.5	BLI Results For Linear Transformation fastText Embeddings	46
Table 5.6	BLI Results For Linear Transformation word2vec Embeddings	46
Table 5.7	BLI Results For Comparison for Iterative Dim. Reduction	49
Table 5.8	BLI Results For NewsFirst data in all steps	49

# 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	vii
1 Introduction	1
1.1 Background	1
1.2 Research Problem	2
1.3 Research Objectives	3
1.4 Contributions	3
1.5 Publications	4
2 Literature Survey	5
2.1 Overview	5
2.2 Vector representation of words	5
2.3 Embedding Models	7
2.3.1 Word2Vec Model	8
2.3.2 FastText Model	9
2.4 Cross-lingual Alignment of Word Vectors	9
2.5 Post-Processing Embedding Spaces	11
2.6 Bilingual Lexical Induction	12
2.6.1 Count-based Vector Space Models	12
2.6.2 Inducing Joint Cross-lingual Embedding Models	14
2.6.3 Projection Based or Mapping Approaches	14
2.6.4 VecMap Model	15
2.7 Post-Processing Embedding Spaces	19
2.8 Summary	20



3	METHODOLOGY	22
3.1	Overview	22
3.2	Model Selection	24
3.2.1	VecMap Model	24
3.2.2	InstaMap Model	24
3.2.3	ClassyMap Model	25
3.2.4	Summary	26
3.3	Extensive analysis on BLI in low resource language pairs	26
3.3.1	Size and the nature of monolingual data	26
3.3.2	Type of The Evaluation Dictionary	28
3.4	Post-processing the Embedding Spaces	29
3.4.1	Linear transformation	29
3.4.2	Dimensionality reduction	30
3.4.3	Linear transformation with dimensionality reduction	31
3.5	Iterative Dimensionality Increment With VecMap	32
3.6	Modified VecMap Model With Pre-processed Embeddings	33
4	EXPERIMENTS	35
4.1	Experimental Setup	35
4.2	Data	35
4.2.1	Corpora and Embeddings	35
4.2.2	Evaluation dictionaries	36
4.3	Experiments	37
4.3.1	Comprehensive analysis	37
4.3.2	Post-processing pretrained embeddings	38
4.3.3	Post-processing with modified VecMap model	39
5	RESULTS AND DISCUSSION	40
5.1	Comprehensive Analysis	40
5.1.1	Pre-trained fastText embeddings	40
5.1.2	Combined News Data	41
5.1.3	Separate news sources	42
5.1.4	Combined news data based on writing styles	44

5.2	Post-processing Pre Trained Embeddings	45
5.2.1	Linear transformation	45
5.2.2	Effective dimensionality reduction	47
5.2.3	Linear transformation + Effective dimensionality reduction	47
5.2.4	Improved VecMap model	48
5.2.5	Post-processing pretrained embeddings+Improved VecMap model	49
5.3	Summary	50
6	CONCLUSION AND FUTURE WORK	52
	References	54