

BILINGUAL LEXICON INDUCTION FOR THE SINHALA-ENGLISH LANGUAGE PAIR

Kasun Wickramasinghe

229405N

Master of Science in Computer Science

Department of Computer Science & Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

March 2024

BILINGUAL LEXICON INDUCTION FOR THE SINHALA-ENGLISH LANGUAGE PAIR

Kasun Wickramasinghe

229405N

Thesis submitted in partial fulfillment of the requirements for the degree
Master of Science in Computer Science

Department of Computer Science & Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

March 2024

DECLARATION

I declare that this is my own work and this Thesis 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: 2024/06/28

The supervisor should certify the Thesis with the following declaration.

The above candidate has carried out research for the Master of Science in Computer Science Thesis under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr. Nisansa de Silva

Signature of the Supervisor:

Date: 29/06/2024

DEDICATION

This research is dedicated to the research community to use the findings here and take the scientific research further forward for a better world.

ACKNOWLEDGEMENT

- I would like to express my special gratitude to my supervisor Dr. Nisansa de Silva for guiding me throughout the research.
- Then I would like to extend my gratitude to all the staff members of the Department of Computer Science and Engineering for conducting lectures and exams and for all the major and minor support given to make the MSc program a success.
- Next, I would like to thank all the staff of the University of Moratuwa who support making this MSc program a success.
- My colleague MSc students for all the help and support.
- Then my special gratitude goes to Emojot (Pvt) Ltd for providing the GCP GPU access for a part of the research experiments.
- The panel of Examiners for evaluating our effort.
- Finally, I acknowledge the research community and the open-source community worldwide for helping to take forward scientific research.

ABSTRACT

Sans a dwindling number of monolingual embedding studies originating predominantly from the low-resource domains, it is evident that multilingual embedding has become the de facto choice due to its adaptability to the usage of code-mixed languages, granting the ability to process multilingual documents in a language-agnostic manner yet monolingual embedding alignment techniques are used in low-resource languages like Sinhala. Our main focus here is to improve the Sinhala word embedding alignment. Here in this research, we experiment with the available monolingual embedding alignment techniques to have the best Sinhala-English embedding alignment that have been achieved so far. For that, first we introduce a large-scale Sinhala-English word-level parallel dictionary that facilitates any word-level cross-lingual tasks. Next, we align Sinhala and English embedding spaces using the available embedding alignment techniques. We identify that Bilingual Lexicon Induction does not measure the true degree of alignment in some cases and we propose solutions for them. We propose a novel stem-based BLI technique to evaluate two aligned embedding spaces which take into account the inflected nature of languages as opposed to the prevalent word-based BLI techniques. Further, we introduce a vocabulary pruning technique that is more informative in showing the degree of the alignment, especially performing BLI on multilingual embedding models. We extend our experiments to 8 other languages as well and prove the validity of our methods. In addition to that using all the languages we conduct a comparative study about the quality of the aligned monolingual embeddings, multilingual embeddings, and hybrid-aligned embeddings. We have published two conference papers during the research so far and we have released all the resources in GitHub ¹² to open-source usage for the research community.

Keywords: word embedding alignment, multilingualism, bilingual lexicon induction, BLI, alignment dictionaries, language-agnostic processing, multilingual embedding, inflected languages, measure alignment, Sinhala

¹<https://github.com/kasunw22/sinhala-para-dict/tree/main>

²<https://github.com/kasunw22/sinhala-word-embedding-alignment/tree/main>

TABLE OF CONTENTS

Declaration of the Candidate & Supervisor	i
Dedication	ii
Acknowledgement	iii
Abstract	iv
Table of Contents	v
List of Figures	viii
List of Tables	ix
List of Abbreviations	ix
List of Appendices	xi
1 Introduction	1
2 Related Work	4
2.1 Parallel Datasets	4
2.1.1 Introduction	4
2.1.2 Parallel Datasets with Sinhala Language	4
2.2 Embedding Generation	5
2.2.1 Vectorized Representations for Words	5
2.2.2 Word Embedding Generation	5
2.2.3 Contextual Embeddings	6
2.2.4 Sentence Embeddings	6
2.2.5 Multilingual Embeddings	7
2.3 Embedding Alignment	7
2.3.1 Word Embedding Alignment Techniques	7
2.3.2 English-Sinhala Embedding Alignment	10
2.3.3 Alignment Datasets	10
2.3.4 Aligned Global Embeddings vs Multilingual Contextualized Embeddings	10
2.3.5 Cross-Lingual Embedding Evaluation Techniques	11

2.3.6	BLI Benchmarking Datasets	11
3	Publications of the Research	12
3.1	Sinhala-English Parallel Word Dictionary Dataset	12
3.2	Sinhala-English Word Embedding Alignment: Introducing Datasets and Benchmark for a Low Resource Language	12
3.3	Aligned or Multilingual: A Comparative Study in Word Embedding Paradigms (under review)	12
4	Sinhala-English Parallel Dictionary Dataset	13
4.1	Introduction	13
4.2	Available Datasets for Sinhala	13
4.3	Dataset Creation	14
4.3.1	Overview	14
4.3.2	Creation Process	15
4.3.3	Dataset Statistics	18
4.3.4	Dataset Evaluation	19
4.3.5	Results and Discussion	20
5	Traditional Alignment Techniques for Sinhala	24
5.1	Methodology	24
5.1.1	Alignment Evaluation	24
5.2	Alignment Dataset Creation	24
5.2.1	Pointwise Mutual Information Criterion	24
5.2.2	Conditional Probability Product	25
5.2.3	Using an Available Dataset	26
5.2.4	Dataset Statistics	26
5.3	Studied Alignment Techniques	28
5.3.1	Orthogonal Procrustes Analysis	28
5.3.2	RCSLS	28
5.3.3	VecMap	29
5.3.4	Contrastive Alignment	30
5.4	Experiments	30
5.4.1	Dataset Comparison	31

5.4.2	Alignment Results	31
5.4.3	Impact of Alignment Dataset Size	32
5.5	Discussion	33
5.5.1	Impact of the embedding model size	33
5.5.2	Quality of the alignment dataset	34
6	Study Contextual Embeddings for Sinhala	36
6.1	Methodology	36
6.1.1	Multilingual Models for Aligned Word Embeddings	36
6.1.2	Aligned Monolingual Models vs Multilingual Models	36
6.1.3	Effect of Inflection	36
6.1.4	Pruning Vocabularies	37
6.1.5	Effect of the Language Family	37
6.2	Experiments	38
6.3	Discussion	41
7	Findings	44
7.1	Large Scale Sinhala-English Parallel Dictionary Dataset	44
7.2	Multilingual Embedding Alignment	44
7.2.1	Conditional Probability-based Alignment Dataset Creation Method	44
7.2.2	Sinhala-English Embedding Alignment	44
7.2.3	Multilingual Embeddings vs Embedding Alignment Study	45
7.2.4	Studies Done for Other Languages	45
7.3	Improvements to BLI	45
7.3.1	Vocabulary Pruning	45
7.3.2	Stem-based BLI	46
8	Discussion and Conclusion	47
	References	48
	Appendix A Extended Evaluation: LaBSE vs LASER	60
	Appendix B Extended Evaluation: All Methods	61
	Appendix C Graphical Representations of Results	64
	Appendix D Computation Costs	70

LIST OF FIGURES

Figure	Description	Page
Figure 4.1	Dictionary 1 Creation Pipeline (For V1 of the Datasets)	16
Figure 4.2	Dictionary 2 and Dictionary 3 Creation Pipeline (For V1 of the Datasets)	17
Figure 4.3	Example of the Quantitative Evaluation Score	22
Figure 5.1	Top-k Retrieval distribution for RCSL alignment. (The numbers indicate how many pairs in the test set are retrieved in $En \rightarrow Si$ and $Si \rightarrow En$ directions with corresponding top-k values)	35
Figure 6.1	Effect of Inflection for Retrieval Accuracy	40
Figure 6.2	Effect of Vocabulary Pruning for Retrieval Accuracy	40
Figure C.1	Top-1 CSLS Retrieval Accuracy comparison for RCSLS vs VecMap vs C1	64
Figure C.2	Top-5 CSLS Retrieval Accuracy comparison for RCSLS vs VecMap vs C1	64
Figure C.3	Top-10 CSLS Retrieval Accuracy comparison for RCSLS vs VecMap vs C1	65
Figure C.4	Top-1 CSLS Retrieval Accuracy comparison for All Alignment Techniques	65
Figure C.5	Top-5 CSLS Retrieval Accuracy comparison for All Alignment Techniques	65
Figure C.6	Top-10 CSLS Retrieval Accuracy comparison for All Alignment Techniques	66
Figure C.7	Top-1, Top-5 and Top-10 CSLS Retrieval Accuracy for RCSLS	66
Figure C.8	Top-1, Top-5 and Top-10 CSLS Retrieval Accuracy for VecMap	66
Figure C.9	Top-1, Top-5 and Top-10 CSLS Retrieval Accuracy for C1	67
Figure C.10	Top-1, Top-5 and Top-10 CSLS Retrieval Accuracy for LaBSE	67
Figure C.11	Top-1, Top-5 and Top-10 CSLS Retrieval Accuracy for C2	67
Figure C.12	Top-1 NN vs CSLS Retrieval Accuracy for RCSLS	68
Figure C.13	Top-1 NN vs CSLS Retrieval Accuracy for VecMap	68
Figure C.14	Top-1 NN vs CSLS Retrieval Accuracy for C1	68
Figure C.15	Top-1 NN vs CSLS Retrieval Accuracy for LaBSE	69
Figure C.16	Top-1 NN vs CSLS Retrieval Accuracy for C2	69

LIST OF TABLES

Table	Description	Page
Table 4.1	Dataset Statistics	19
Table 4.2	Experimental Results on Wikimedia dataset	20
Table 4.3	Experimental Results on TED2020 dataset	21
Table 5.1	Dataset Statistics: Statistics of the alignment datasets we have experimented with	27
Table 5.2	En-Si Procrustes Embedding Alignment Results of cc-Fasttext embeddings on different datasets	31
Table 5.3	English-Sinhala word translation average precisions (@1, @5, @10) for Traditional Alignment Techniques	32
Table 5.4	Extended Comparison of Traditional Alignment Techniques with Other Languages	32
Table 5.5	Si→En Embedding Alignment Results with previous alignment work	33
Table 5.6	En→Si Procrustes Embedding Alignment Results with different dataset sizes	33
Table 6.1	Word translation precision of different multilingual models	39
Table 6.2	Impact of the inflections of Sinhala language for alignment evaluation	39
Table 6.3	En-Si BLI Score Improvements by Stem-based BLI	39
Table 6.4	The Effect of Vocabulary Pruning for BLI score	40
Table 6.5	LaBSE Embedding-based BLI Score Improvements after Vocabulary Pruning	41
Table 6.6	Top-1 BLI Retrieval accuracy for different Retrieval Criteria	41
Table 6.7	English-Sinhala word translation average precisions (@1, @5, @10) for All Alignment Techniques	42
Table A.1	Word translation average precision for LaBSE and LASER embeddings (@1, @5, @10)	60
Table B.1	Word translation average precision for all alignment techniques (@1, @5, @10)	62

LIST OF ABBREVIATIONS

Abbreviation	Description
BLI	Bilingual Lexicon Induction
CBOW	Continuous Bag-of-Words Model
CSLS	Cross-domain similarity local scaling
ELMo	Embeddings from Language Models
IC	Inverse Consultation
mMLM	Multilingual Masked Language Modelling
MT	Machine translation
NLP	Natural Language Processing
qWP	Quantized Wasserstein Procrustes
SVD	Singular Value Decomposition
TLM	Translation Language Modeling
WP	Wasserstein Procrustes

LIST OF APPENDICES

Appendix	Description	Page
Appendix -A	Extended Evaluation: LaBSE vs LASER	60
Appendix -B	Extended Evaluation: All Methods	61
Appendix -C	Graphical Representations of Results	64
Appendix -D	Computation Costs	70