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 [1][2] 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

LIST OF FIGURES

LIST OF TABLES

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