CROSS-LINGUAL DOCUMENT CLUSTERING FOR SINHALA, TAMIL, AND ENGLISH USING PRE-TRAINED MULTILINGUAL LANGUAGE MODELS

Malavarayar Vijayanandan Vithulan

(209390R)

Degree of Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

July 2022

CROSS-LINGUAL DOCUMENT CLUSTERING FOR SINHALA, TAMIL, AND ENGLISH USING PRE-TRAINED MULTILINGUAL LANGUAGE MODELS

Malavarayar Vijayanandan Vithulan

(209390R)

Thesis report submitted in partial fulfilment of the requirements for the degree Master of Science in Computer Science specialisation in Data Science.

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

July 2022

DECLARATION

I declare that this is my work. 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. To the best of my knowledge and belief, it does not contain any previously published material written by another person except where the acknowledgment is made in the text.

Also, I hereby grant the University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/dissertation, in whole or in print, electronic, or another medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 05/05/2022

The above candidate has researched the Master's thesis Dissertation under my supervision.

Name of the supervisor: Dr. Surangika Ranathunga

Signature of the supervisor: Date:

ABSTRACT

Organising text articles into groups or clusters is known as document clustering. Documents that belong to a cluster are about the same subject. Document embeddings should be in the same embedding space for the cross-lingual document clustering, i.e., similar documents should have similar vectors. Obtaining document embedding for Tamil and Sinhala is feasible using models like Word2Vec or FastText, however, these embeddings are language specific, i.e., these will not be in the same vector space. Therefore, one cannot cluster documents across the languages using the language specific models. Pre-trained multilingual language models such as mBERT, XLM-R were introduced to solve this problem by transferring the knowledge from high resource languages to low resource languages.

This research is conducted to cluster Tamil, Sinhala and English news articles using XLM-R models. An adequate amount of collected documents were clustered, and the clustering techniques and performance were evaluated. This research produces a new baseline for cross-lingual clustering of Tamil, Sinhala, and English documents.

Keywords: Cross-lingual document clustering, Multilingual language models, XLM-R, mBERT, LASER, Knowledge distillation

ACKNOWLEDGEMENT

I want to express my deep and sincere gratitude to Dr. Surangika Ranathunga for guiding me in finding an exciting research topic and for continued support and encouragement. Her supervision immensely helped me in setting goals and engaging in the study.

I want to express my most incredible gratitude to the Department of Computer Science and Engineering, the University of Moratuwa, for providing the support to overcome this effort. Last but not least, my heartfelt gratitude goes to my parents and friends who supported me throughout this endeavour.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Description
NLP	Natural Language Processing
NER	Named-entity recognition
MLLM	Multilingual Language Models
mBERT	Multilingual BERT