QUESTION ANSWERING SYSTEM FOR SINHALA LANGUAGE

Hewa Lunuwilage Amila Viraj Mahinda

208537T

Master of Science in Artificial Intelligence

Department of Computational Mathematics
Faculty of Information Technology

University of Moratuwa
Sri Lanka

April 2023

QUESTION ANSWERING SYSTEM FOR SINHALA LANGUAGE

Hewa Lunuwilage Amila Viraj Mahinda

208537T

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree Master of Science in Artificial Intelligence

Department of Computational Mathematics
Faculty of Information Technology

University of Moratuwa
Sri Lanka

April 2023

DECLARATION

I declare that this thesis is the result of my independent effort and has not been previously presented in any form for the purpose of obtaining another degree or diploma from any university or tertiary institution. Acknowledgment has been provided within the text for any information or findings derived from the published or unpublished works of others, and a comprehensive list of references is included.

Signature: Date: 03 July 2023

The above candidate has carried out research for the Master's thesis/dissertation under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr. KASN Sumathipala

Signature of the Supervisor: Date: 10 July 2023

DEDICATION

I dedicate this thesis to the individuals who have provided unwavering support and encouragement throughout my academic journey. It is with deep gratitude that I express my heartfelt appreciation to those who have played a pivotal role in my journey.

Foremost, I extend my sincerest appreciation to my beloved family, whose unwavering support and belief in my abilities have been the driving force behind my achievements. Your sacrifices and steadfast encouragement have been the cornerstone of my success, and I am forever grateful for your unwavering presence in my life.

I would also like to express my profound gratitude to my supervisor and the esteemed academic staff. Their guidance, mentorship, and profound expertise have been instrumental in shaping my research and nurturing my intellectual growth. Their invaluable feedback and constructive criticism have challenged me to reach new heights and have honed my skills in ways I never thought possible. I am truly indebted to their unwavering dedication and commitment to my academic development.

To my all-dear friends, I am incredibly grateful for the camaraderie, laughter, and unyielding support you have bestowed upon me. Your companionship has provided solace during challenging times and has infused my academic journey with joy and positivity. Your unwavering encouragement has been a constant source of motivation, reminding me of the significance of perseverance and friendship.

To all those mentioned above and to those whose influence may go unspoken but is cherished nonetheless, I offer my deepest gratitude. Your unwavering support, guidance, and encouragement have propelled me forward and have helped me achieve this significant milestone in my academic career.

May this work not only contribute to the realm of knowledge but also serve as a testament to the collective efforts of those who have accompanied me on this remarkable journey.

ACKNOWLEDGEMENT

I would like to express my deepest gratitude and appreciation to the everyone who have contributed significantly to the completion of this research project. Their unwavering support, guidance, and encouragement have been invaluable, and I am indebted to their contributions.

First, I extend my heartfelt appreciation to Dr. Sagara Sumathipala, Senior Lecturer in the Department of Computational Mathematics at the Faculty of Information Technology, University of Moratuwa. I am grateful for his exceptional mentorship, expert guidance, and continuous support throughout the entire duration of this project. His profound knowledge, insightful feedback, and unwavering commitment to academic excellence have been instrumental in shaping the direction and quality of this research. I am truly grateful for his invaluable contributions.

I would also like to express my sincere gratitude to Dr. Subha Fernando, Senior Lecturer in Computational Mathematics at the Faculty of Information Technology, University of Moratuwa. Her invaluable advice, encouragement, and thoughtful insights have played a pivotal role in helping me achieve my goals. Her expertise in the field and her willingness to share her knowledge have been instrumental in enhancing the depth and rigor of this research. I am truly grateful for her unwavering support and guidance.

Furthermore, I would like to extend my special thanks to my friends and family. Their constant encouragement, understanding, and support have been a source of motivation and strength throughout the challenges encountered during this research journey. Their belief in my abilities and their unwavering support during moments of failure and setbacks have been invaluable. I am deeply grateful for their presence in my life and their unconditional support.

ABSTRACT

Natural Language Processing (NLP) has seen rapid progress in recent years, thanks to the development of advanced models such as BERT (Bidirectional Encoder Representations from Transformers) and their variants. However, the research and application of NLP remain focused on high-resource languages, leaving low-resource languages with limited resources and tools. In this thesis, we present the process of developing a question-answering (QA) system for the Sinhala language, a low-resource language, using the state-of-the-art transformer models. To prepare the Sinhala dataset, we use one of the most popular benchmark QA datasets in English called SQuAD 2.0 dataset and undertake the translation of the SQuAD 2.0 dataset into Sinhala and proceed to pre-process and adapt it to create a SQuAD-like format. I describe the use of large pre-trained language model fine-tuning for the QA task. My experimental results demonstrate the suitability of XLM-RoBERTa-Large for the Sinhala language and highlight the potential of Transformer models to address the challenges faced by low-resource languages in the field of NLP. By evaluating the model using standard QA metrics, such as F1 score and exact match score, it provides insights into the model's performance and compare it with related works. With the experiments, I found that the use of already down-streamed language model for QA task and fine tuning it with custom Sinhala data has outperformed the other models with a significant increasement of the accuracy based on F1 & exact match scores. Finally, I discuss the limitations and suggest potential improvements and future research directions in building cutting-edge NLP models for the low-resource languages, emphasizing the significance of our commitment to empowering underrepresented languages in the NLP domain.

Keywords: Question-Answering, Natural Language Processing, Low-Resource Languages, Sinhala Language, Translation, Large Language Model Fine-Tuning, SQuAD 2.0, Reading Comprehension.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Description
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations for Transformers
EM	Exact Match
GPU	Graphics Processing Unit
NLP	Natural Language Processing
NLU	Natural Language Understanding
QA	Question Answering
SQuAD	Stanford Question Answering Dataset
TAR	Translate Align Retrieve
TPU	Tensor Processing Units
UI	User Interface