

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 Acknowledgment has been provided within the text for any information or findings derived from the published or unpublished works of others, and a comprehensive list of references is included.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 increase 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

Declaration	i
Dedication	ii
Acknowledgement	iii
Abstract	iv
Table of Contents	v
List of Figures	viii
List of Tables	x
List of Abbreviations	xi
Chapter 1	1
Introduction	1
1.1 Prolegomena	1
1.2 Objectives	2
1.3 Background and Motivation	2
1.4 Problem Definition	3
1.5 Question Answering System for Sinhala Language	4
1.6 Structure of the thesis	4
1.7 Summary	5
Chapter 2	7
Literature Review	7
2.1 Introduction	7
2.2 Developments of Question Answering Systems in English	7
2.3 Question Answering for Low Resource Languages	8
2.4 Analysis of Literature	9
2.5 Analysis of Different QA Datasets in English	10
2.6 Analysis of SQuAD-like Dataset in Other Languages	11
2.7 Summary	12
Chapter 3	13
technologies adopted	13
3.1 Introduction	13
3.2 Programming Languages	13
3.3 Natural Language Processing	13
3.3.1 Hugging Face Transformers	13
3.3.2 Simple Transformers	14
3.4 Machine Learning, Deep Learning Frameworks	14

3.4.1	TensorFlow	14
3.4.2	PyTorch.....	14
3.5	Cloud Computing	14
3.6	Language Translation	15
3.7	Data management	16
3.8	User Interface (UI)	16
3.9	Summary	16
Chapter 4		17
Question Answering Systems for Sinhala Language		17
4.1	Introduction	17
4.2	Hypothesis	17
4.3	Inputs	17
4.4	Outputs	18
4.5	Process.....	18
4.6	Users	19
4.7	Summary	19
Chapter 5		20
Design		20
5.1	Introduction	20
5.2	High Level Design of the System.....	20
5.3	Sinhala Data Preparation	21
5.3.1	Sinhala Translation Module	21
5.3.2	Post Processing Module	22
5.4	QA System Development & Inference.....	23
5.4.1	Fine-Tuning Module	23
5.4.2	Inference Module	23
5.5	Summary	24
Chapter 6		25
Implementation		25
6.1	Introduction	25
6.2	Data Collection & Preparation	25
6.2.1	SQuAD Data Format.....	25
6.2.2	Data Preprocessing.....	27
6.2.3	Sinhala Translation	29
6.2.4	Find Sinhala Answers	30
6.2.5	Format the Dataset	32
6.3	QA Model Development	33

6.3.1	Split the Dataset into Train & Test	34
6.3.2	Selecting Language Models	35
6.3.3	Model Training	36
6.4	QA Model Inference.....	41
6.5	Web Interface	42
6.6	Summary	45
Chapter 7		46
Evaluation		46
7.1	Introduction	46
7.2	Experimental Dataset.....	46
7.3	Evaluation Metrics	46
7.3.1	F1 Score	46
7.3.2	Exact Match (EM).....	47
7.4	Model Performance Evaluations	47
7.4.1	Fine-tuned Multilingual BERT Base Cased Model	48
7.4.2	Fine-tuned Multilingual Distil-BERT Base Cased Model	49
7.4.3	Fine-tuned XLM RoBERTa Large Squad Model	50
7.4.4	Comparison of Model Performance Evaluations	51
7.4	Discussion of Findings	52
7.4	Summary	52
Chapter 8		54
Conclusion & further work		54
References		56

LIST OF FIGURES

Figure	Description	Page
Figure 5.1	High Level System Design	20
Figure 5.2	Preparation Process of Sinhala SQuAD Dataset	21
Figure 5.3	QA System Development & Production Overview	23
Figure 6.1	Format of SQuAD 2.0 Data	26
Figure 6.2	Sample of SQuAD 2.0 Dataset	26
Figure 6.3	Check Relevant Data inside SQuAD Dataset	27
Figure 6.4	Sample Output of Relevant Data	27
Figure 6.5	Function to Extract Data as Document Level from SQuAD 2.0	28
Figure 6.6	Sample Output from the Function with Extracted Data	28
Figure 6.7	Adding Special Characters into the Context	28
Figure 6.8	Sample Output for Context after Adding Special Characters	28
Figure 6.9	Translation of Data into Sinhala Language	29
Figure 6.10	Translated Context Data Sample	29
Figure 6.11	Script for Question Translation	30
Figure 6.12	Script to Filter Data	31
Figure 6.13	Function to Extract the Sinhala Answer	31
Figure 6.14	Sample Output of all the Prepared Data	32
Figure 6.15	Script to Format the Dataset	32
Figure 6.16	Sample Data Example from Prepared Sinhala SQuAD Dataset	33
Figure 6.17	Information of the Selected Dataset	33
Figure 6.18	Split Data into Train & Test	34
Figure 6.19	Make Train Data into SQuAD Format	34
Figure 6.20	Make Test Data into SQuAD Format	35
Figure 6.21	BERT Model Fine Tuning	37
Figure 6.22	Distil-BERT Model Fine Tuning	40
Figure 6.23	RoBERTa-SQuAD Model Fine Tuning	41

Figure 6.24	Model Inference	41
Figure 6.25	List of Possible Answers Predicted by the Model with Confidence Scores	42
Figure 6.26	Back-end script for the web application	43
Figure 6.27	Front-end script for the web application Inference	44
Figure 6.28	Web Application of Sinhala-QA-Extractor Solution	45
Figure 7.1	Evaluate Script using F1 Score and EM Metrics	48
Figure 7.2	Training Loss - BERT	48
Figure 7.3	Evaluation - BERT	49
Figure 7.4	F1 & EM Scores - BERT	49
Figure 7.5	Training Loss – Distil-BERT	50
Figure 7.6	Evaluation – Distil-BERT	50
Figure 7.7	F1 & EM Scores – Distil-BERT	50
Figure 7.8	Training Loss – RoBERTa	51
Figure 7.9	Evaluation – RoBERTa	51
Figure 7.10	F1 & EM Scores – RoBERTa	51

LIST OF TABLES

Table	Description	Page
Table 2.1	Analysis of Literature	9
Table 2.2	Analysis of QA English Datasets	10
Table 2.3	Analysis of SQuAD-like Datasets in Other Languages	11
Table 7.1	Fine-Tuned Model Performance Comparison	52

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