DOMAIN SPECIFIC QUESTION AND ANSWER GENERATION IN TAMIL

Rubika Murugathas

199358H

M.Sc. in Computer Science

Department of Computer Science and Engineering

University of Moratuwa

SriLanka

July 2022

DOMAIN SPECIFIC QUESTION AND ANSWER GENERATION IN TAMIL

Rubika Murugathas

199358H

This dissertation submitted in partial fulfillment of the requirements for the Degree of MSc in Computer Science specializing in Software Architecture

Department of Computer Science and Engineering

University of Moratuwa

SriLanka

July 2022

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

UOM Verified Signature 19.10.2022

Rubika Murugathas Date

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

UOM Verified Signature 19.10.2022

Dr. Uthayasanker Thayasivam Date

ABSTRACT

Automatic Question-Answer generation is a challenging task in natural language processing. A system developed is capable of automatically generating questions and answers from history related text content in Tamil language input by user. The system processes the input text using various NLP techniques and generates questions and answers. The system has four modules namely, Preprocessing module, Rule-based module, Named Entity Recognition (NER) module, Question Answer Generator(QAG) module. Regex patterns and gazetteers are used in rule-based module and machine learning approach is used for NER module. NER module uses Conditional Random Fields (CRF) classifier built with features suitable for the domain and language. Dataset is collected from history textbooks and 23k word tokens are tagged using IOB2 format. Novel entity tag set specific to history domain are tagged. NLP techniques such as Sentence tokenization, POS tagging, Stemming, Unicode conversion uses existing python libraries. Features suitable for the domain and language selected are experimented with multiple combination. POS tag, stem word, gazetteer and clue words are features that contributes more for the performance. The best feature combination produced micro averaged Precision, Recall, F1-score of 87.9%, 67.1% and 76.1% respectively and accuracy of 89.6% on the test dataset. The NER module produced a better results despite the domain & language related challenges. Questions are formed using grammatical and defined rules from the named entities identified from rule-based and NER module. An affix stripping algorithm implemented to find the inflection suffix. A history text from Wikipedia is evaluated by 16 native Tamil speakers under categories such as undergraduates, graduates and experts. According to the evaluation results, 62.22% of total generated questions are grammatically correct and meaningful questions. Questions generated from Rule-based module produces better results compared to NER module.

Keywords: question and answer generation, tamil, NER, CRF, history, domain specific

ACKNOWLEDGEMENTS

My profound gratitude goes to Dr. Uthayasanker Thayasivam, my supervisor for the knowledge, supervision, advice and guidance provided with his expertise, throughout in making this thesis a success.

My appreciation goes to my family for the motivation and support they have provided for my studies. I also would like to thank my colleagues in the MSc batch and at my workplace for the help and support provided in managing my research work.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Description
BERT	Bidirectional Encoder Representations from Transformers
BiGRU	Bidirectional Gated Recurrent Unit
BiLSTM	Bidirectional Long Short Term Memory
CNN	Convolutional Neural Networks
CRF	Conditional Random Field
GRU	Gated Recurrent Unit
LSTM	Long Short Term Memory
MA	Morphological Analysis
NE	Named Entity
NER	Named Entity Recognition
NLP	Natural Language Processing
NQG	Neural Question Generation
POS	Part-of-Speech
QA	Question and Answer
QAG	Question and Answer Generation
RNN	Recurrent Neural Networks

LIST OF APPENDICES