

DEVELOPING A RETRIEVAL-BASED TAMIL LANGUAGE CHATBOT FOR CLOSED DOMAIN

Kumaran Kugathanan

198097X

Thesis/Dissertation submitted in partial fulfilment of the requirements for
the degree Master of Science in Computer Science and Engineering

Department of Computer Science & Engineering

University of Moratuwa

Sri Lanka

August 2023

DECLARATION

I declare that this is my own work and this dissertation 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 dissertation, 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).

Signature:

Date: 14/08/2023

The above candidate has carried out research for the Masters thesis/Dissertation under my supervision.

Signature of the Supervisor:

Date: 15/08/2023

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Dr. Uthayasanker Thayasivam, my supervisor, for his invaluable advice and guidance throughout this project. I am grateful for his willingness to make himself available whenever I needed his assistance. Without his support and encouragement, I could not have completed this project successfully.

Additionally, I would like to thank members of my progress review committee Prof. Sanath Jayasena and Dr. Charith Chitraranjan, for their insightful feedback and guidance, which were extremely helpful to me.

I am also grateful to the entire lecturers and research students at the DataSEARCH Research Centre for their assistance, feedback and the resources they provided to carry out the project.

Lastly, I would like to say thank you to my family and friends for their unfailing support.

ABSTRACT

Chatbots are conversational systems that interact with humans via natural language. Frequently, it is used to respond to user queries and provide them with the information they need. To build a highly functional chatbot, a good corpus and a variety of language-related resources are required. Since Tamil is a low-resource language those resources are not available for Tamil. Additionally, since Tamil is also a morphologically rich language, high inflexion and free word order pose key challenges to Tamil chatbots. Due to all the above reasons, it is evident that developing an effective End-to-End chat system is challenging even for a closed domain.

This study introduces a novel method for building a chatbot in Tamil by leveraging a dataset extracted from Tamil banking website's FAQ sections and extending it to encompass the language's morphological complexity and rich inflectional structure. Intent is assigned to each query, and a multiclass intent classifier is developed to classify user intent. The CNN-based classifier demonstrated the highest performance, achieving an accuracy of 98.72%.

While previous works on short-text classification in Tamil focused only on a few classes and used a very large dataset, our method produced a superior accuracy of over 98% using a small number of per-class examples even when there are 56 classes and additional challenges like class imbalance problem in the data. This shows our approach is better than any other approach for short text classification in Tamil.

The major contribution of this research is the generation of the first-ever chat dataset for Tamil. Our research is the first of its kind in Tamil to show how an efficient context-less chatbot can be built using short text classification. Although this project is done for the Tamil language and for the Banking domain, this approach can be applied to other low-resourced languages and domains as well.

LIST OF FIGURES

Figure 1: Problems with using human agents	9
Figure 2: Interest in chatbot over time	10
Figure 3: Open Domain Chatbot	11
Figure 4: Closed Domain Chatbot	11
Figure 5: Goal oriented chatbot	12
Figure 6: Non-Goal oriented Chatbot	12
Figure 7 - Level 1 Chatbot. Adapted from [19]	12
Figure 8 - Level 2 Chatbot. Adapted from [19]	12
Figure 9 - Level 3 Chatbot. Adapted from [19]	13
Figure 10: Evolution of Chat technology	19
Figure 11: Pattern matching system	19
Figure 12: Eliza chatbot	20
Figure 13 : Parsing	21
Figure 14 : Parse Tree	21
Figure 15 : Markov Chain model	22
Figure 16 : AIML code block 1	24
Figure 17 : AIML code block 2	25
Figure 18 : AIML code block 3	25
Figure 19 : Chat script	26
Figure 20 : Poongkuzhali Chatbot	30
Figure 21 : Cody Chatbot	31
Figure 22 : Machan Chatbot	31
Figure 23 : Methodology	36
Figure 24 : Class imbalance	39
Figure 25 : Developed Chat Web Application	41

LIST OF TABLES

Table 1: Sample English to Tamil Obligue Translation	32
Table 2 : Sample sentences from dataset	40
Table 3 : Model Performance	43
Table 4 : Chatbot testing results	45

LIST OF ABBREVIATIONS

FAQ	Frequently Asked Questions
NLU	Natural Language Understanding
NLG	Natural Language Generation
AIML	Artificial Intelligence Markup Language
POS	Part-of-speech
NER	Named-Entity Recognizer
BOW	Bag of Words
TF-IDF	Term frequency–inverse document frequency
SMOTE	Synthetic Minority Oversampling Technique

TABLE OF CONTENTS

DECLARATION	1
ACKNOWLEDGEMENT	2
ABSTRACT	3
LIST OF FIGURES	4
LIST OF TABLES	5
LIST OF ABBREVIATIONS	6
1. INTRODUCTION	9
1.1 Type of chatbot	10
1.2 Problem in Low-Resource Chatbots	13
1.3 Problems with Tamil Chatbot	14
1.4 Problem Statement	14
1.5 Motivation	15
1.6 Objective	15
1.7 Research Contributions	16
1.8 Organisation	17
1.9 Chapter Summary	17
2. EVOLUTION OF CONVERSATIONAL SYSTEMS	19
2.1 Pattern matching	19
2.2 Parsing	20
2.3 Markov Chain Models	22
2.4 Ontologies	23
2.5 Artificial Intelligence Markup Language (AIML)	24
2.6 Chat Script	26
2.7 Machine Learning	27
2.8 Chapter Summary	28
3. LITERATURE REVIEW	29
3.1 English Language Chat Systems	29
3.2 Low Resource Chat Systems	29
3.3 Tamil Language Chat Systems	30
3.4 Techniques used for dataset creation	32
3.5 Short Text Classification	33
3.6 Evaluation Metrics	35
3.7 Chapter Summary	35
4. METHODOLOGY	36
4.1 Dataset creation	36
4.2 Intent classifier	37
4.3 End-to-End chatbot development	38
4.4 Chapter Summary	38
5. EXPERIMENTS	39

5.1 Dataset	39
5.2 Machine Learning Models	41
5.3 Chat Application	41
5.4 Chapter Summary	42
6. RESULTS	43
6.1 Model Performance	43
6.2 Chatbot Performance	44
6.3 Chapter Summary	46
7. DISCUSSION	47
7.1 Model Performance	47
7.2 Chabot Performance	47
7.3 Limitations	48
7.3.1 Limitations in Generated Dataset	48
7.3.2 Limitations in the Approach	48
7.3.3 Limitations in the Testing Strategy	49
7.4 Chapter Summary	50
8. CONCLUSION	51
8.1 Future work	51
8.1.1 Develop a better dataset	51
8.1.2 Create Multiple Responses	51
8.1.3 Multi-level classification	51
8.1.4 Confidence Score for Classification	52
8.2 Chapter Summary	53
REFERENCES	55