



A Hybrid Machine Learning and Rule-Based Approach for a Sinhala Natural Language Interface to Database

M. R. M. Mahdi
Reg. No.: MS24010344

A THESIS
SUBMITTED TO
SRI LANKA INSTITUTE OF INFORMATION TECHNOLOGY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

December 2025

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Ms. Jenny Krishara

Approved for MSc. Research Project:

MSc in IT Programme Co-ordinator, SLIIT

Approved for MSc:

Head of Graduate Studies, FoC, SLIIT

DECLARATION

This is to certify that the work is entirely my own and not of any other person, unless explicitly acknowledged (including citation of published and unpublished sources). The work has not previously been submitted in any form to the Sri Lanka Institute of Information Technology or to any other institution for assessment for any other purpose.



Sign:

M. R. M. Mahdi

Date:03/11/2025.....

ABSTRACT

A Hybrid Machine Learning and Rule-Based Approach for a Sinhala Natural Language Interface to Database

MSc. in Information Technology

December 2025

The objective of Natural Language Interfaces to Databases (NLIDBs), is to provide users with an intuitive way to get at their data; users can ask questions of their relational databases using natural language, instead of using a formal query language like SQL. While there have been significant advancements in developing NLIDs for high resource languages such as English, support for low resource and morphologically rich languages such as Sinhala continues to be limited. Most existing Sinhala NLIDBs have employed rule-based approaches, these have limitations in terms of adapting to new conditions and scaling.

In this paper we propose a hybrid approach to developing Sinhala NLIDBs that combines rule-based logic with statistical methods to address the current limitations of Sinhala NLIDBs. Our focus will be on a single student table and supporting the basic SQL operations. We will employ a combination of core linguistic preprocessing techniques (tokenization, stemming, POS-tagging) along with a grammar driven query parser that is specifically designed to accommodate the unique structure of Sinhala.

We will use a manually annotated dataset of 800 Sinhala-SQL query pairs to improve our model's ability to identify semantic elements through Named Entity Recognition (NER). Furthermore, we will employ an intent classifier to guide the SQL generation process, enabling us to correctly understand a variety of natural language queries. Our hybrid architecture seeks to achieve a balance between the precision of rule-based systems and the flexibility of statistical systems; providing both interpretability and generalizability. In addition to improving the accessibility of databases for users who speak Sinhala, this research provides a foundation for developing future multilingual and multidomain NLIDBs for low resource languages.

ACKNOWLEDGEMENT

First and foremost, I am deeply grateful to **Almighty Allah** for granting me the strength, wisdom, and patience to successfully complete this research journey. Without His guidance and blessings, none of this would have been possible.

I would like to extend my sincere thanks to **Ms. Jenny Krishara**, my academic mentor at the **Sri Lanka Institute of Information Technology**, whose insightful guidance, critical feedback, and unwavering support greatly helped to me to complete this research successfully.

I am also profoundly thankful to my **parents and family**, who have been my greatest source of strength and inspiration throughout this journey. Their unconditional love, prayers, and moral support kept me motivated even during the most difficult times.

To my **friends and peers**, thank you for the camaraderie, brainstorming sessions, and late-night conversations that made this academic experience both enriching and enjoyable.

Lastly, I wish to acknowledge my **work colleagues**, who supported my research efforts amidst professional responsibilities and encouraged me to strive for balance and excellence.

To all those who contributed to this endeavor in various ways, your presence and support have meant more than words can express.

TABLE OF CONTENTS

DECLARATION	ii
ABSTRACT	iii
ACKNOWLEDGEMENT	iv
TABLE OF CONTENTS	v
List of Figures	x
List of Tables	x
Chapter 1 Introduction	1
1.1 Chapter Overview	1
1.2 Problem Domain	2
1.3 Problem Definition	2
1.4 Research Motivation	3
1.5 Existing Work	4
1.6 Research Gap	5
1.7 Research Challenges	6
1.8 Research Aim	7
1.9 Research Question	8
1.10 Research Objectives	8
1.11 Scope	9
1.11.1 In Scope	10
1.11.2 Out of Scope	10
1.11.3 Prototype Features	11
1.12 Contribution to the Body of Knowledge	11
1.12.1 Contribution to Software Engineering/ Computer Science	12
1.12.2 Contribution to the Domain	13
1.13 Resource Requirements	14
1.13.1 Hardware Resource Requirements	14
1.13.2 Software Resource Requirements	14
1.13.3 Skills Requirements	15
Chapter 2 Literature Review	16
2.1 Chapter Overview	16
2.2 Existing Work on NLIDB Systems	16
2.2.1 Early Rule-Based Systems	17
2.2.2 Shift Toward Statistical and Machine Learning Methods	17
2.2.3 Hybrid and Knowledge-Driven Approaches	18
2.2.4 Evaluation and Benchmarking	18

2.2.5 Summary	18
2.3 NLIDB in Non-English and Low-Resource Languages	19
2.4 Sinhala NLIDB Systems	20
2.4.1 SNLIDB (2020)	20
2.4.2 SNLIDB v2 (2021).....	21
2.4.3 SinSQLFinder (2021).....	21
2.4.4 Summary	22
2.5 Justification for a Hybrid Approach.....	22
Chapter 3 Methodology	25
3.1 Chapter Overview	25
3.2 Research Methodology	25
3.2.1 Research Philosophy and Approach	25
3.2.2 Research Strategy.....	26
3.2.3 Research Design.....	26
3.2.4 Data Collection and Preparation	27
3.2.5 Research Techniques and Tools.....	27
3.2.6 Research Validation and Reliability	28
3.2.7 Ethical Considerations	28
3.3 Development Methodology	28
3.3.1 Overview.....	28
3.3.2 Iterative Prototype Model	29
3.3.3 Hybrid Development Approach	30
3.3.4 Version Control and Collaboration	30
3.3.5 Data Annotation Workflow.....	30
3.3.6 Model Training and Integration Process	31
3.3.7 Testing and Feedback Cycle	31
3.3.8 Tools and Technologies	32
3.3.9 Summary	32
3.4 Project Management Methodology.....	33
3.4.1 Overview.....	33
3.4.2 Project Management Approach.....	33
3.4.3 Project Phases	34
3.4.4 Risk Management	35
3.4.5 Milestone Tracking and Deliverables	35
3.4.6 Summary	36
3.5 Evaluation Methodology.....	36
3.6 Rich Picture.....	37

3.7 Stakeholder Analysis	37
3.8 Selection of Requirement Elicitation Techniques.....	38
3.8.1 Self-Evaluation	38
3.8.2 Literature Review.....	38
3.8.3 Questionnaires.....	38
3.9 Discussion of Results from Elicitation.....	40
3.10 Summary of Findings.....	41
3.11 Context Diagram.....	41
3.12 Use Case Diagram and Description	43
3.13 Functional Requirements	48
3.14 Non-Functional Requirements	48
3.15 Chapter Summary	49
Chapter 4 Social, Legal, Ethical and Professional Issues (SLEP)	50
4.1 Chapter Overview	50
4.2 Social Considerations.....	50
4.3 Legal Considerations	50
4.4 Ethical Considerations	51
4.5 Professional Issues	52
4.6 Risk and Mitigation Summary	52
4.7 Chapter Summary	53
Chapter 5 System Architecture and Design	54
5.1 System Overview	54
5.2 System Architecture Diagram.....	55
5.2.1 Stage 1: Pre-Processing - Syntax Analysis	56
5.2.2 Stage 2: Machine Learning Driven Stage	56
5.2.3 Stage 3: Post-Processing - Semantic Analysis.....	57
5.2.4 Architectural Summary	57
5.3 Input Processing Pipeline.....	57
5.3.1 Tokenization	58
5.3.2 Stemming	58
5.3.3 Part-of-Speech (POS) Tagging	58
5.3.4 Lexical Analysis.....	59
5.3.5 Summary	60
5.4 NER and Intent Classification.....	60
5.4.1 Workflow Integration.....	60
5.4.2 Dataset Overview	61
5.4.3 Named Entity Recognition (NER)	61

5.4.4 Intent Classification	64
5.5 Semantic Analysis and SQL Construction.....	66
5.5.1 Structured Rule Application.....	66
5.5.2 Dynamic Condition Handling	67
5.5.3 Template Flexibility and Fallbacks.....	68
5.5.4 Post-Processing and Validation	68
5.5.5 Summary	68
Chapter 6 Implementation.....	70
6.1 Development Environment	70
6.1.1 Programming Language and Libraries.....	70
6.1.2 Labeling and Annotation Tools.....	70
6.1.3 Stemming and Syntax Tools	71
6.1.4 Hardware and Execution Environment.....	71
6.2 Dataset Preparation	71
6.3 Named Entity Annotation Process	72
6.3.1 Entity Labels Used.....	73
6.3.2 Annotation Workflow	73
6.4 Intent Annotation Process	74
6.4.1 Intent Classes Defined	74
6.4.2 Annotation Methodology	75
6.4.3 Use in the System.....	75
6.5 Conversion Formats	76
6.5.1 Named Entity Data: CoNLL Format.....	76
6.5.2 Intent Data: CSV Format	77
6.5.3 Why Conversion Matters	77
6.6 Model Training Details	78
6.6.1 Named Entity Recognition (NER) Training	78
6.6.2 Intent Classifier Training	79
6.6.3 Model Packaging.....	79
6.7 Evaluation Plan	80
6.7.1 Module-Level Evaluation	80
6.7.2 System-Level Evaluation	81
Chapter 7 Results and Discussion.....	82
7.1 Chapter Overview	82
7.2 Named Entity Recognition (NER) Performance.....	82
7.3 Intent Classification Performance.....	83
7.4 End-to-End SQL Generation Accuracy	83

7.5 Error Reduction Cycle	84
7.6 Discussion	84
7.7 Comparative Benchmarking with Existing Sinhala NLIDB Systems.....	85
7.8 Stakeholder and Expert Evaluation.....	86
7.9 Functional Testing Summary	87
7.10 Achievement of Research Objectives	88
7.11 Limitations	88
7.12 Discussion	89
7.13 Chapter Summary	89
Chapter 8 Conclusion and Future Work	91
8.1 Chapter Overview	91
8.2 Summary of Research Achievements	91
8.3 Achievement of Research Aims and Objectives.....	92
8.4 Problems and Challenges Faced	93
8.5 Limitations of the Research	93
8.6 Future Enhancements	94
8.7 Concluding Remarks.....	95
References.....	96
Appendix.....	97
Appendix 1: UI Screenshots	97
Appendix 2: Black-Box Testing Results.....	98

List of Figures

Figure 3.1 - Rich Picture of the Sinhala NLIDB Hybrid System.....	37
Figure 3.2 - Context Diagram of the Sinhala NLIDB System	42
Figure 3.3 - Use Case Diagram of the Sinhala NLIDB System.....	43
Figure 5.1 - Architecture of the proposed Sinhala NLIDB system.....	55
Figure 5.2 - Pre-processing output of each token in a Sinhala sentence	59

List of Tables

Table 3.1 - The research design combined quantitative and qualitative elements	26
Table 3.2 - Iterative Development Cycles and Key Outcomes	29
Table 3.3 - Tools and Libraries Used in the Implementation	32
Table 3.4 - Research Phases and Key Deliverables	34
Table 3.5 - Research Risks and Corresponding Mitigation Strategies	35
Table 3.6 - Stakeholder Analysis and Objectives	37
Table 3.7 - Questionnaire Summary and Interpretation (n = 50)	39
Table 3.8 - Functional and Non-Functional Requirements Overview	41
Table 3.9 - Use Case 1: Enter Sinhala NLQ	43
Table 3.10 - Use Case 2: Generate SQL	44
Table 3.11 - Use Case 3: Display NLP Insights	45
Table 3.12 - Use Case 4: Execute SQL.....	46
Table 3.13 - Use Case 5: Display Result	47
Table 3.14 - Functional Requirements of the Sinhala NLIDB Hybrid System.....	48
Table 3.15 - Non-Functional Requirements of the Sinhala NLIDB Hybrid System	48
Table 4.1 - Identified Risks and Mitigation Strategies	52
Table 5.1 - NER Model Training and Configuration Parameters	62
Table 5.2 - Intent Classifier Model Training and Configuration Parameters.....	64
Table 7.1 - NER Model Performance Metrics	82
Table 7.2 - Intent Classifier Model Performance Metrics.....	83
Table 7.3 - End-to-End SQL Generation Accuracy Results	83
Table 8.1 - Comparative Benchmark of Sinhala NLIDB Systems	85
Table 8.2 - User Evaluation Summary.....	86
Table 8.3 - Evaluation of Research Objects.....	88
Table 9.1 - Summary of Research Objectives and Outcomes.....	92