NLIDB Converter for Customer Relationship Index

G. O. Wijayasekara

149236 L

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for the partial fulfillment of the requirements of the Degree of Master of Science in Information Technology.

May 2017

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Gayanee Wijayasekara ………………………………
Name of Student Signature of Student

………………………………
Date

S.C. Premaratne ………………………………
Name of Supervisor Signature of Supervisor

....
Date

Acknowledgement

I take this opportunity to express my heartfelt gratitude to my supervisor, Mr. S. C. Premaratne, Senior Lecturer at the University of Moratuwa, Sri Lanka, for guiding me throughout the research project without whose support and direction this would not have been possible.

And a special heartfelt thanks goes out to my loving family for being the pillars of strength that they are at all times.

I would also like to thank my colleagues for the valuable feedback and support they extended during the course of the research.

Abstract

Information plays a significant role in our day to day life. Information affects how we look at things and make decisions. Therefore, accurate information must be readily available. With technology reaching new heights, in modern day systems and applications, the main source for information storage is an underlying database system. Retrieving information from a database management system requires a specific expert skill set which predominantly includes knowledge on Structured Query Language and domain specific knowledge. In recent times there is a rising demand for non-expert users to be able to directly extract information from an underlying database management system. For industry specific Customer Relationship Management applications like the Customer Relationship Index used in large organizations like the Brandix Group this is seen as both a challenge and an opportunity to be explored. Implementation of a Natural Language Interface to the underlying database of the Customer Relationship Index allows a typical user to retrieve the required information from the underlying database using natural language like English without prior programming or technical knowledge. For above reasons the NLIDB converter for the Customer Relationship Index is introduced. This application takes the users query to be run against the underlying database in natural language and returns the corresponding T-SQL statement which can then be run against the database to extract results. The NLIDB converter has shown that it can successfully and efficiently convert questions in natural language to the corresponding T-SQL statement with an accuracy rate of more than 80% for two types of output T-SQL statement formats for the Customer Relationship Index database.

Contents

1 Introduction 1

1.1 Prolegomena 1

1.2 Background and Motivation 2

1.3 Problem in Brief 4

1.4 Hypothesis 6

1.5 Objectives 6

1.6 Proposed Solution 7

1.7 Structure of the Thesis 7

1.8 Summary 7

2 Challenges in Natural Language Interfaces for Databases 8

2.1 Introduction 8

2.2 Early Developments in Natural Language Interfaces to Databases 8

2.3 Approaches Used for NLIDB 11

2.3.1 Symbolic Approach (Rule Based Approach) 11

2.3.2 Empirical Approach (Corpus Based Approach) 12

2.3.3 Connectionist Approach (Using Neural Network) 12

2.4 Techniques Used to Implement NLIDB 13

2.4.1 Pattern Matching Techniques 13

2.4.2 Syntax Based Techniques 13

2.4.3 Semantic Grammar Techniques 14

2.5 Recent Developments in NLIDB 14

2.6 Research Problem 15

2.7 Summary 16

3 Technology 17

3.1 Introduction 17

3.2 State of the art in natural language processing 17

3.2.1 Stanford CoreNLP 18
3.2.2 Stanford Parser for .NET 18
3.3 Microsoft .NET Framework 19
3.4 Console Applications 19
3.5 Three Tier Architecture 19
3.6 Microsoft Entity Framework 21
3.7 Microsoft SQL Server 21
3.8 Nugget Packages 22
3.9 Summary 22
4 Approach 23
4.1 Introduction 23
4.2 Hypothesis 23
4.3 Input 23
4.4 Output 24
4.5 Process 25
4.6 Features 27
4.7 Users 28
4.8 Summary 28
5 Design 29
5.1 Introduction 29
5.2 Top Level Architecture 29
5.3 The User Interface 30
5.4 The Linguistic Component 31
5.5 Business Logic Module 35
5.6 Data Access Component 35
5.7 NLIDB Database Component 35
5.8 Summary 37

6 Implementation 38

6.1 Introduction 38

6.2 Implementation Essentials for the NLIDB converter for Customer Relationship Index 38

6.2.1 Softwares Used 38

6.2.2 Hardware Requirements..39

6.2.3 Frameworks Used 39

6.2.4 Languages Used for Implementation 39

6.3 Implementation of the NLIDB Database 39

6.4 Implementation of the NLIDB converter for Customer Relationship Index 42

6.4.1 Class Diagram..43

6.4.2 Implementation of the Data Access Component 44

6.4.3 Implementation of the Business Logic Module 45

6.4.4 Implementation of the Linguistic Component 63

6.4.5 Implementation of the User Interfaces..66

6.4.6 Assumptions and Constraints during Implementation..67

6.5 Summary 67

7 NLIDB converter for Customer Relationship in Practice..68

7.1 Introduction 68

7.2 Initial Evaluation of the NLIDB converter for Customer Relationship Index 68

7.3 Scenario 1..71

7.3.1 Scenario Description..71

7.3.2 Input String 71

7.3.3 Arriving at the output..71

7.3.4 Output 74

7.3.5 Results..76

7.4 Scenario 2..77

7.4.1 Scenario Description 77
7.4.2 Input String 77
7.4.3 Arriving at the output 77
7.4.4 Output 81
7.4.5 Results 82
7.5 Scenario 3 82
7.5.1 Scenario Description 82
7.5.2 Input String 82
7.5.3 Arriving at the output 83
7.6 Output 84
7.7 Summary 86
8 Evaluation 87
8.1 Introduction 87
8.2 Participants 87
8.3 Test Cases 89
8.4 Testing Setup 98
8.5 Data Collection 98
8.6 Data Analysis 100
8.6.1 Algorithm Analysis 101
8.6.2 Usability and Performance Analysis 102
8.7 Summary 103
9 Conclusion and Future Work 104
9.1 Introduction 104
9.2 Overall Conclusion 104
9.3 Objective-wise Conclusions 106
9.4 Achievements 108
9.5 Limitations and Future work 108

9.6 Summary 109

10 References 110

List of Figures

Figure 3.1 Layers of the three tier architecture 20

Figure 4.1 Wireframe design for NLIDB converter for Customer Relationship Index – Inputs 24

Figure 4.2 Wireframe design for NLIDB converter for Customer Relationship Index – Outputs 25

Figure 4.3 Process diagram of the NLIDB converter for Customer Relationship Index 27

Figure 5.1 Top level architecture of NLIDB converter for Customer Relationship Index 29

Figure 5.2 Wireframe diagram for NLIDB converter for Customer Relationship Index - Inputs 31

Figure 5.3 Wireframe diagram for NLIDB converter for Customer Relationship Index - Output 31

Figure 6.1 Database diagram of NLIDB 40

Figure 6.2 The three tier architecture in the NLIDB converter for Customer Relationship Index implementation 42

Figure 6.3 Class diagram of NLIDB converter for Customer Relationship Index 43

Figure 6.4 Variables and Methods in NLIDBDAC 45

Figure 6.5 Flow chart - NLIDB converter for Customer Relationship Index 48

Figure 6.6 NLIDB Converter - Code segments - library imports 53

Figure 6.7 NLIDB Converter - Code segments - handling existing questions 54

Figure 6.8 NLIDB Converter - Code segments - handling new questions and configuring Stanford CoreNLP 55

Figure 6.9 NLIDB Converter - Code segments - Extracting token details 56

Figure 6.10 NLIDB Converter - Code segments - Extracting token details 57

Figure 6.11 NLIDB Converter - Code segments - SQL node mapping for Tables and Views 58

Figure 6.12 NLIDB Converter - Code segments - SQL node mapping for Columns and T-SQL statement generation 59

Figure 6.13 NLIDB Converter - Code segments - Generate SQL statement 59

Figure 6.14 Variables and Methods in the NLIDBBAL 61

Figure 6.15 Variables and Methods in the NLIDBLogger 62

Figure 6.16 Variables and Methods in ErrorLogger..63
Figure 6.17 Adding the Stanford CoreNLP to the project...63
Figure 6.18Adding the Stanford Parser to the project 63
Figure 6.19 Code segment - Setting up the path for the Stanford CoreNLP models..64
Figure 6.20 Code segment - Configuring annotators in Stanford CoreNLP................64
Figure 6.21 Code segment - Sentence splitting in Stanford CoreNLP 65
Figure 6.22 Code segment - Tokenizing in Stanford CoreNLP...................................65
Figure 6.23 Code segment - Extracting token properties in Stanford CoreNLP 65
Figure 6.24 Code Segment - Extracting dependencies in Stanford Parser 66
Figure 7.1 Schema of the table containing customer details 69
Figure 7.2 Screenshot of the details in the table BuyerDetail......................................70
Figure 7.3 Initial Evaluation – Scenario 1 – Inputs 71
Figure 7.4 Initial Evaluation -Scenario 1 - Outputs...75
Figure 7.5 Initial Evaluation - Scenario 1 – Results 76
Figure 7.6 Initial Evaluation - Scenario 2 - Inputs...77
Figure 7.7 Initial Evaluation - Scenario 2 – Outputs 81
Figure 7.8 - Initial Evaluation - Scenario 2 – Results..82
Figure 7.9 Initial Evaluation - Scenario 3 – Inputs..82
Figure 7.10 Initial Evaluation - Scenario 3 – Outputs 85
Figure 8.1 Data Collection in NLIDB converter for Customer Relationship Index....99
Figure 8.2 Summary of Evaluation Results 100

List of Tables

Table 1 Summary of recent developments........15
Table 2 Penn Treebank tags........34
Table 3 Lexicon table in NLIDB database........36
Table 4 The description and its role of the tables in the NLIDB database........41
Table 5 Schema explanation of BuyerDetail........71
Table 6 Tokens in Scenario 1........72
Table 7 Natural language rules for NNT in scenario 1........72
Table 8 Token dependencies for tokens in scenario 1........73
Table 9 Natural language rules for NNC in scenario 1........74
Table 10 Tokens mapped to SQL nodes in scenario 1........74
Table 11 Tokens in Scenario 2........77
Table 12 Natural language rules for NNT in scenario 2........78
Table 13 Token dependencies for tokens in scenario 2........79
Table 14 Natural language rules for NNC in scenario 2........80
Table 15 Tokens mapped to SQL nodes in scenario 1........80
Table 16 Tokens in Scenario 3........83
Table 17 Natural language rules for NNT in scenario 3........84
Table 18 Test case 1........90
Table 19 Test case 2........91
Table 20 Test case 3........92
Table 21 Test case 4........93
Table 22 Test case 5........94
Table 23 Test case 6........95
Table 24 Test case 7........96
Table 25 Test case 8........97
Table 26: Summary of Evaluation Results........100
Table 27 Feedback summary questionnaire: G1 and G2........102
Table 28 Feedback summary questionnaire: G3 and G4........103