INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

QueryLift: A Rule Based Approach to Parse Complex Natural Language Queries to SQL

A dissertation by

Mr. Buwaneka Ranmal

W1800734 - 20200006

Supervised by

Mr. Mohamed Cassim Farook

Submitted in partial fulfillment of the requirements for the MSc in Advanced Software Engineering degree at the University of Westminster.

September 2023

ABSTRACT

Over the years, relational database management systems (RDBMs) have been in use among both technical and non-technical professionals. However, it is apparent that the query languages that are used to extract data from RDBMs require extensive knowledge. Hence, ample research has been conducted on natural language interfaces for query languages such as SQL. Nevertheless, previous research contains some limitations including limited support for unseen databases, limited accuracy for complex natural language expressions and complexity of generating constants in SQL. It is evident that the limitation of parsing complex natural language expressions to SQL queries has rarely been addressed in the existing literature.

A rule-based approach is applied in this research to overcome the limited accuracy of translating complex natural language expressions to SQL. Thereby, a unique set of rules is defined to identify table names, column names, conditions, table joins, aggregate operators, GROUP BY and ORDER BY scenarios that are associated with a given natural language expression. Furthermore, the rules are derived by applying dependency parsing and regular expressions.

The accuracy of QueryLift was evaluated using a self-composed dataset which has yielded 84% accuracy. Spider dataset was used to conduct competitive benchmarking and ensure cross domain adaptability. The system, QueryLift was compared with NaLIR and Templar which are state of the art rule-based systems. Thereby, QueryLift has revealed 4.1% accuracy which is higher than the accuracy of both NaLIR and Templar.

Keywords: Natural Language Processing (NLP), Natural Language to SQL, Dependency Parsing, SQL,

Subject Descriptors:

- Computing Methodologies → Artificial Intelligence → Natural Language Processing → Machine Translation
- Information Systems → Data Management Systems → Query Languages → Relational Database Query Languages → Structured Query Language