INFORMATICS INSTITUTE OF TECHNOLOGY

INFORMATICS INSTITUTE OF TECHNOLOGY
In Collaboration with
UNIVERSITY OF WESTMINSTER

NatDB – A text-to-SQL system

A dissertation by
Mr. Thungu Hapuarachchi
W1789989 / 2019788

Supervised by
Mr. Sudharshana Welihindha

Submitted in partial fulfilment of the requirements for the BEng in Software Engineering degree at the University of Westminster.

May 2023

ABSTRACT

The field of database management often presents challenges to individuals lacking the technical expertise necessary for constructing and executing SQL queries. Implementing text-to-SQL systems, which translate user-friendly natural language queries into structured SQL commands, is a viable solution to this problem.

This thesis investigates the creation of a novel text-to-SQL system that addresses the deficiencies identified in the current landscape of such translation systems. The primary objective of this research is to address situations in which lengthy or complex natural language descriptions cannot be accurately converted to SQL queries. This study presents a novel combination of a pre-trained RoBERTa model and a relation-aware transformer network in order to bridge this divide. This merger seeks to improve translation accuracy by combining the rich linguistic understanding of RoBERTa with the relation-aware transformer's ability to comprehend the database schema.

The system was developed using advanced deep learning techniques and an architecture resembling transformer networks. The model was trained to highlight key elements within the natural language query and semantically correlate them with the corresponding SQL command. The system was exhaustively trained and evaluated on a the most popular of benchmark datasets – Spider, exhibiting impressive performance across the defined spectrum of query complexities. However, the study also reveals potential limitations in coping with extremely complex queries and large databases, highlighting areas for further investigation and development. This study contributes to the expanding field of natural language processing and its applications for improving human-computer interactions with its findings.

Keywords: Text-to-SQL, Natural Language Processing, Deep Learning, RoBERTa, Relation-Aware-Transformer, Database Management, SQL queries

Subject Descriptors:
Computing methodologies -> Artificial Intelligence -> Natural Language Processing
Information Systems -> Database Management -> Query languages (SQL)
Information Systems -> Information Interfaces and Presentation -> User Interfaces -> Natural Language Interfaces