TEXT-TO-SQL GENERATION USING SCHEMA ITEM CLASSIFIER AND ENCODER-DECODER ARCHITECTURE

Mohamed Ramiz Aadhil Rushdy

(219175X)

Degree of Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

July 2023

TEXT-TO-SQL GENERATION USING SCHEMA ITEM CLASSIFIER AND ENCODER-DECODER ARCHITECTURE

Mohamed Ramiz Aadhil Rushdy

(219175X)

Thesis submitted in partial fulfilment of the requirements for the degree Master of Science in Computer Science specialization in Data Science, Engineering and Analytics

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

July 2023

DECLARATION

I hereby affirm that the content presented in this dissertation is solely my original work. It does not incorporate any material from previously submitted works for a degree or diploma at any other university or institute of higher learning. Furthermore, to the best of my knowledge and belief, this dissertation does not contain any previously published or written material by another individual, except where explicit acknowledgement is provided within the text.

Furthermore, I hereby authorize the University of Moratuwa to reproduce and distribute my dissertation, in whole or in part, using a variety of media formats, including print, electronic, or any other medium. Moreover, I maintain the privilege to utilize the content of my dissertation, either in its entirety or partially, for future endeavors such as the creation of articles or books.

Signature: ……………… Date: ………17/07/2023

Name: M.R. Aadhil Rushdy

I certify that the declaration above by the candidate is true to the best of my knowledge and he has researched the Master's thesis dissertation under my supervision.

Signature of the supervisor: ………………. Date: ……………

Name: Dr. Uthayasanker Thayasivam

Senior Lecturer, Dept of Computer Science and Engineering, University of Moratuwa, Moratuwa 10400, Sri Lanka.

Acknowledgement

I would like to express my sincere thanks and gratitude to my supervisor Dr. Uthayasanker Thayasivam for leading me throughout the research and assisting me unconditionally to gather all necessary resources to complete this thesis. And I am grateful for his invaluable assistance in the research work with the necessary expertise, tools, guidance, supervision and helpful suggestions.

I am eternally grateful to my parents and sister for their unconditional love and unwavering support throughout my M.Sc. journey. I would also like to express my heartfelt appreciation to my beloved wife, whose constant encouragement and support were instrumental in my success.

Finally, I would like to extend my deepest gratitude to all those individuals who have directly and indirectly guided and supported me in the completion of this research. Their invaluable contributions have played a pivotal role in our research's successful completion.

Abstract

The objective of the text-to-SQL task is to convert natural language queries into SQL queries. However, the presence of extensive text-to-SQL datasets across multiple domains, such as Spider, introduces the challenge of effectively generalizing to unseen data. Existing semantic parsing models have struggled to achieve notable performance improvements on these cross-domain datasets. As a result, recent advancements have focused on leveraging pre-trained language models to address this issue and enhance performance in text-to-SQL tasks. These approaches represent the latest and most promising attempts to tackle the challenges associated with generalization and performance improvement in this field. I proposed an approach to evaluate and use the Seq2Seq model by giving the most relevant schema items as the input to the encoder and to generate accurate and valid cross-domain SQL queries using the decoder by understanding the skeleton of the target SQL query. The proposed approach is evaluated using Spider dataset which is a well-known dataset for text-to-sql task and able to get promising results where the Exact Match accuracy and Execution accuracy has been boosted to 72.7% and 80.2% respectively compared to other best related approaches.

Keywords: Text-to-SQL, Seq2Seq model, BERT, RoBERTa, T5-Base

Table of Contents

1 INTRODUCTION 1

1.1 RESEARCH PROBLEM 2

1.1.1 Challenges in Natural Language 2

1.1.2 Challenges in SQL 3

1.2 RESEARCH OBJECTIVES 3

2 LITERATURE REVIEW 4

2.1 Schema Linking 5

2.1.1 Candidate Discovery 5

2.1.2 Candidate Matching 5

2.2 Approaches for Natural Language representation 6

2.2.1 Global Vectors (GloVe) 6

2.2.2 Bidirectional Encoder Representation from Transformers (BERT) 6

2.2.3 Bidirectional and Auto-Regressive Transformers (BART) 7

2.2.4 Text-to-Text Transfer Transformer (T5) 7

2.3 Approaches for Encoding 7

2.4 Approaches for Decoding 13

2.5 Error Analysis 15

2.5.1 Column Prediction 15

2.5.2 Nested Query 15

2.5.3 Operator 15

3 METHODOLOGY 16

3.1 Data Collection 16

3.1.1 WikiSQL 16

3.1.2 ATIS 17

3.1.3 Spider [28] 17

3.1.4 CoSQL [39] 17

3.1.5 SParC [3] 18

3.2 Overall Architecture 18

3.2.1 Schema Items Classifier 19

3.2.1.1 RoBERTa model for encoding 19

3.2.1.2 Bi-LSTM as the classifier 19

3.2.2 T5-Base Seq2Seq model 19

3.2.2.1 Encoder 19

3.2.2.2 Decoder 19

3.2.3 Implementation 20

4 Experiment and Evaluation 21

4.1 Evaluation Metrics 21

4.2 Baseline Models 21

4.3 Results on Spider [28] dataset 21

5 Discussion 23

6 Conclusion 24

7 REFERENCES 25

LIST OF FIGURES

Figure	Description	Page
Fig. 1.1	Text-to-SQL problem	2
Fig. 2.1	General Steps in a Text-to-SQL system	4
Fig. 2.2	Overview of the neural model employed for synthesizing an intermediate representation known as SemQL [5]	8
Fig. 2.3	An example of Erosion [6]	8
Fig. 2.4	An example of Shuffle [6]	9
Fig. 2.5	Overview of the IGSQL model [9]	10
Fig. 2.6	The BERT-Based deep model	11
Fig. 2.7	Encoding step proposed in ValueNet [19]	11
Fig. 2.8	A layer in the RAT Schema Encoder [10]	12
Fig. 2.9	Encoder proposed in Grappa including the approach of pre-training [31]	12
Fig. 2.10	The overall model architecture proposed in LGESQL [30]	13
Fig. 2.11	Execution Guided Decoding	14
Fig. 2.12	Illustration of Tree structured architecture in RAT-SQL	14

decoder [10]

Fig. 2.13 Illustration of PICARD and beam search for constrained decoding [26] 15

Fig. 3.1 Proposed Architecture Overview 18

LIST OF TABLES

Table Description Page

Table 1 Example from Spider [28] dataset 20

Table 2 Comparison of accuracies for older approaches 22

Table 3 Comparison of accuracies new state of the art approaches 22

Table 4 Example correct predictions from the spider [28] dataset 23

LIST OF ABBREVIATIONS

Abbreviation	Description
RNN	Recurrent Neural Networks
Bi-LSTM	Bi-directional Long short-term Memory networks
NLP	Natural Language Processing