SHAH AND ANCHOR KUTCHHI ENGINEERING COLLEGE, CHEMBUR
Department of Computer Engineering
University of Mumbai
2020-2021

Mini Project

on

Generating SQL query from Natural Language using NLP

Under Guidance of

Prof. Deepshikha Chaturvedi

Team Members:
Rakesh Choudhary(TE3-08)
Satyam Gupta(TE3-14)
Yogesh Hole(TE3-15)

Table of contents

Problem Statement

- To create a system which converts Natural Language Query to SQL query by taking input as text format.
- Input should be processed to generate SQL query and execute it to provide output.

Purpose

Need for conversion of NL Query to SQL Query

- The user must type in the required SQL Query to access or process data. Users that aren't acquainted with SQL, on the other hand, are unable to retrieve relevant data.
- It is very tedious for non-technical person to get the information from database they need without needing to understand SQL.

Introduction

- Natural Language Processing(NLP)
  - What is NLP?
  - Use of NLP

- System Design
  - System Architecture
  - Features

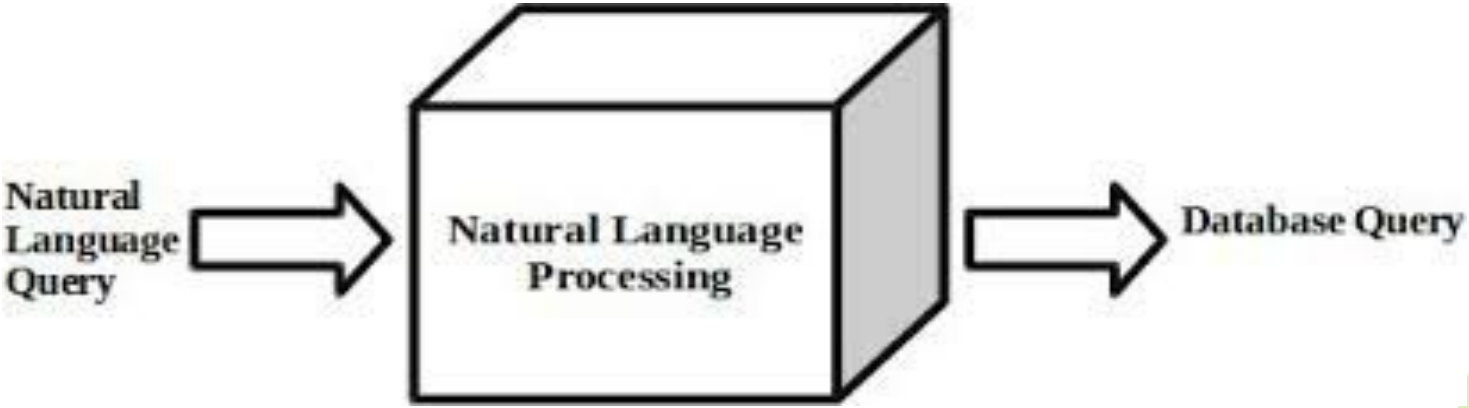
Literature Survey

Sr. No	Author	Title	Methodology	Objective	Limitation
1	M. Uma, V. Sneha, G. Sneha, J. Bhuvana and B. Bharathi, 2019	Formation of SQL from Natural Language Query using NLP	NLP Phase: -Tokenization -Lemmatization -Syntactical Analysis -Semantic Analysis Mapping Phase: - Attribute Identification -Sql Query Formation	Formation of SQL query using Natural Language Processing on train seats and fare datasets.	They have taken input in text form and they have make use of only one table.
2	Alaka Das, Rakesh Chandra Balabantaray, 2019	MyNLIDB: A Natural Language Interface to Database	Input□Pre-Processing□(tagger)□Node generator□SQL generator□Database □Result	To make Natural Language Interface to Database.(i.e. Interface through which we retrieve data from database)	It is made only for simple queries
3	A. kate, S. Kamble, A. Bodhke, M. Joshi, 2018	Conversion of Natural Language Query to SQL Query.	Tokenization , Lexical Analysis , Syntactic Analysis , Semantic Analysis.	To help T&P officer to easily retrieve and manage student data from database.	This system does not provide high accuracy in results.

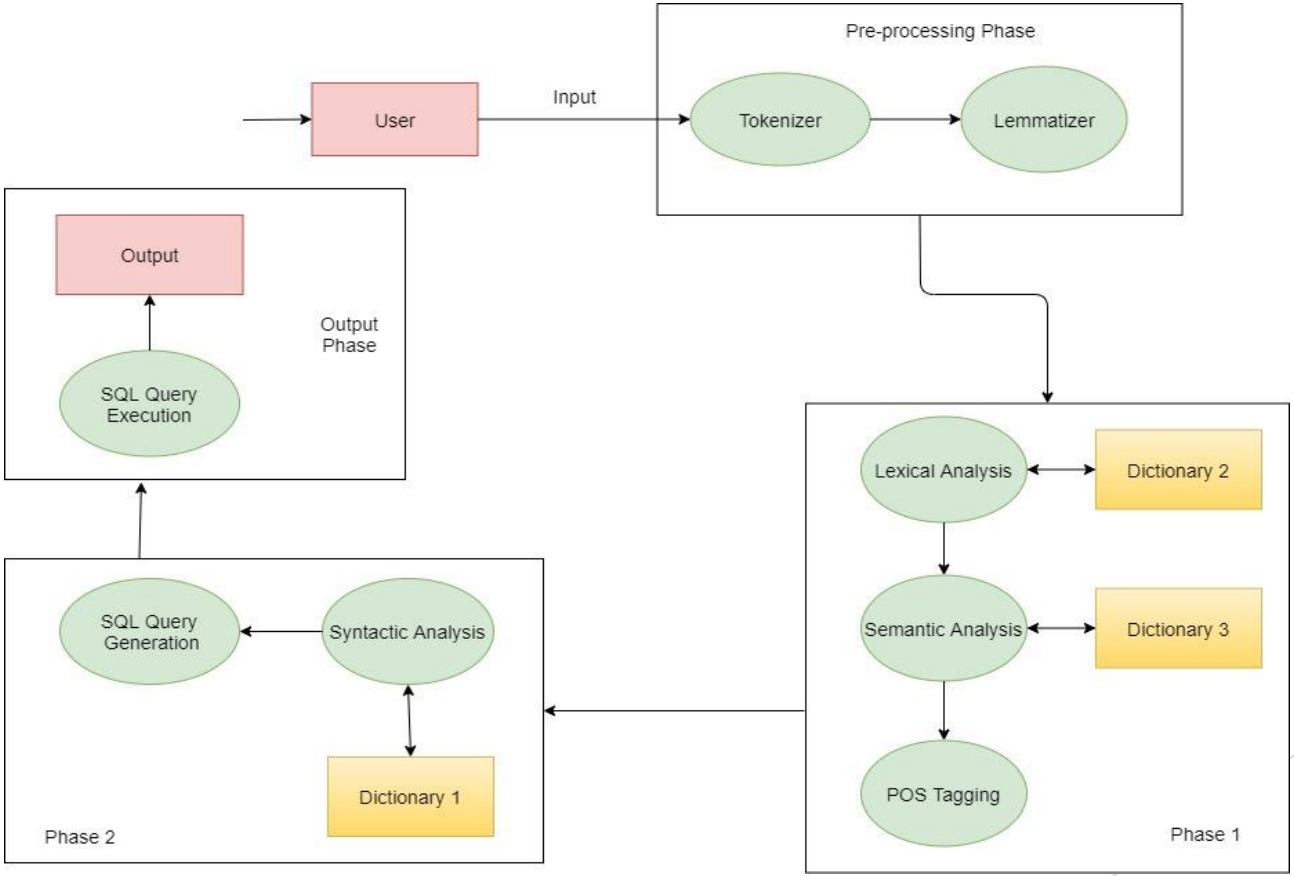
Sr. No	Author	Title	Methodology	Objective	Limitation
4	Tanzim Mahmud, K. M. Azharul Hasan, Mahtab Ahmed, Thwoi Hla Ching Chak, 2015	A Rule Based Approach for NLP Based Query Processing	Design context free grammars word check,removing excess words, tokenization and mapping to CFG rules.	Develop NLID for alumni database using CFG base system	Accuracy is very low i.e 47%
5	N. Sangeeth , R. Rejimoan , 2015	An Intelligent System For Information Extraction From Relational Database Using HMM	Two modules used: Linguistic module, database module	develop NLIDB system based on HMM using GEOQUERY database	Accuracy and Performance can be increased, audio input is not taken
6	Vishal Wudaru, Aruneswara Reddy, Radhika Mamid, 2019	Question Answering on Structured Data using NLIDB Approach	Intermediate query approach Syntactic approach 1.Tokenizer & POS Tagging 2.Stop word remover 3.Dependency Parsing	Proposed system able to handle most queries related to movie domain through database	Focused on limited number of database attributes

Proposed Solution

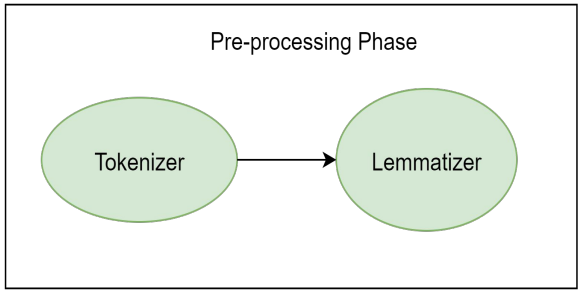
Generating SQL Query from Natural Language Query using NLP



Proposed model



Preprocessing phase



- **Input**

  fetch all the information of the students who have scored more than 70 in 10th
- **Tokenizer**

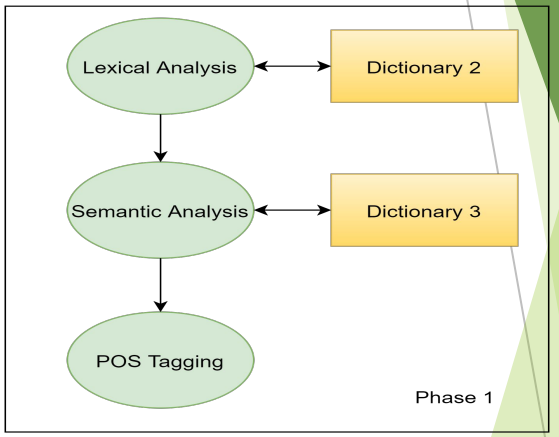
  ['fetch', 'all', 'the', 'information', 'of', 'the', 'students', 'who', 'have', 'scored', 'more', 'than', '70', 'in', '10th’]
- **Lemmatization**

  ['fetch', 'all', 'student', 'who', 'have', 'more', '70', '10th’]

Phase 1

- **Lexical Analysis**

  ['SELECT', '*', 'FROM student', 'WHERE', 'more', '70', 'ssc’]



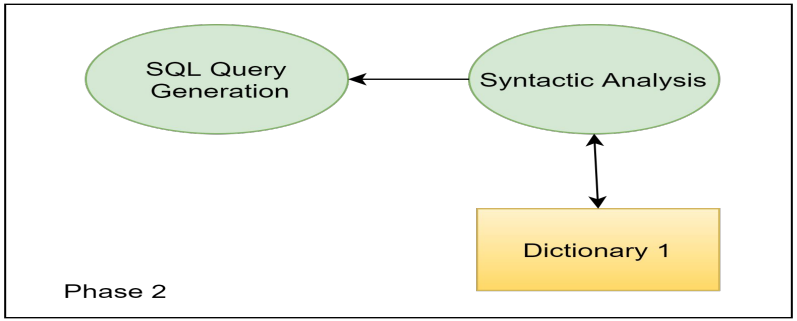
- **Semantic Analysis**

  ['SELECT', '*', 'FROM student', 'WHERE', '>', '70', 'ssc’]

- **POS Tagging**

  [('SELECT', 'NNP'), ('*', 'NNP'), ('FROM student', 'NNP'), ('WHERE', 'NNP'), ('>', 'VBZ'), ('70', 'CD'), ('ssc', 'NN’)]

Phase 2



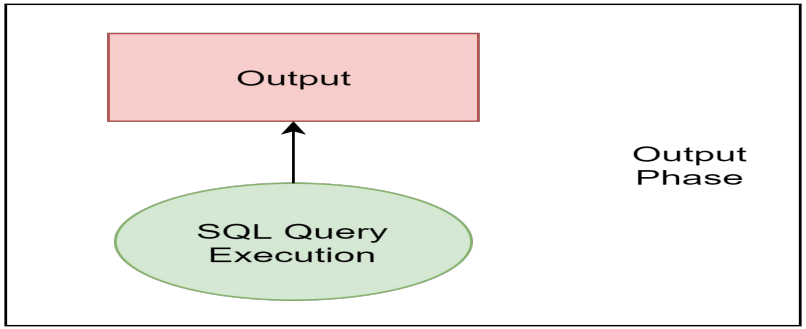
- **Syntactic Analysis**

  ['SELECT'] ['*'] ['FROM student'] ['ssc'] ['>'] ['70'] []

- **SQL Query Generation**

  SELECT * FROM student WHERE ssc > 70

Output Phase



- **SQL Query Execution**

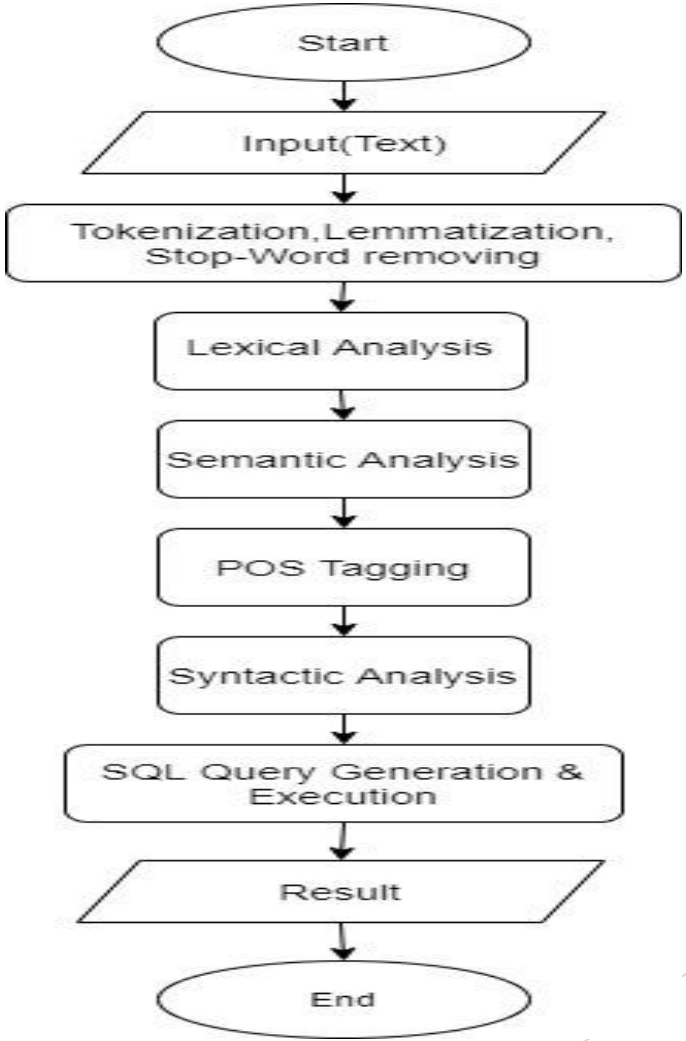
  SQL Query generated will be executed on students database

- **Output**

tk

1	rakesh	80	90	85
2	satyam	85	75	80
3	yogesh	90	70	80
6	simran	95	85	90
8	devesh	80	50	65
9	sharvil	75	65	70
10	shubham	80	80	80
19	rupesh	73	79	76
25	deepak	91	99	95

Block Diagram



Results & Analysis

- This system converts Natural Language Query to SQL Query efficiently and effectively.
- Tokenization, syntactic, and semantic analysis, as well as the use of dictionaries and grammar is used for such analysis is performed using the natural language to SQL query conversion tool.
- Libraries used – MySQL, Tkinter, word_tokenize, wordNetLemmatizer, pos_tag by nltk
- System has achieved accurate results for many natural language queries.

Conclusion & Future scope

- A system has been built successfully which converts Natural Language Query to SQL Query using NLP.
- The given proposed system only works for simple queries.
- It uses only single table, Further work can be done for multiple tables and for executing complex queries.

References

[1] A. kate, S. Kamble, A. Bodhke, M. Joshi, “Conversion of Natural Language Query to SQL Query,” 2018 IEEE Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India,2018.

[2] M. Uma, V. Sneha, G. Sneha, J. Bhuvana and B. Bharathi, "Formation of SQL from Natural Language Query using NLP," 2019 IEEE International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, India, 2019.

[3] Alaka Das, Rakesh Chandra Balabantaray, “MyNLIDB: A Natural Language Interface to Database”, 2019 IEEE International Conference on Information Technology (ICIT), Bhubaneswar, India,2019.

[4] Tanzim Mahmud, K. M. Azharul Hasan, Mahtab Ahmed, Thwoi Hla Ching Chak, “A Rule Based Approach for NLP Based Query Processing ,“ 2015 Khulna University of Engineering & Technology (KUET) Khulna – 9203, Bangladesh,2015 IEEE.

[5] Vishal Wudaru, Nikhil Koditala, Aruneswara Reddy, Radhika Mamid, “Question Answering on Structured Data using NLIDB Approach”, 2018 International Institute of Information Technology, Hyderabad, India, 2018 IEEE.

Appendix

1. Dictionary 1 is used to store the SQL Query in parts according to it clauses.
2. Dictionary 2 is used to check the grammar of the words, so it may be replaced by database words
3. Dictionary 3 is to check the conditions in entered Query
4. ignore_list=['the','record','database','table','information','a','are','is','to','marks', 'mark','of','in','than',"'s",',','.','me','you','us','an','score','scored']

Thank You!