International Journal of Innovative Research in Computerand Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 6, June 2020||

SQL Query Formation for Database System using NLP

Aditya Narhe[1], Chaitanya Mohite[2], Rushikesh Kashid[3], Pratik Tade[4], Santosh Waghmode[5]

U.G. Student, Department of Computer Engineering, JSPM'S Imperial College of Engineering, SPP University, Pune, India[1]

U.G. Student, Department of Computer Engineering, JSPM'S Imperial College of Engineering, SPP University, Pune, India[2]

U.G. Student, Department of Computer Engineering, JSPM'S Imperial College of Engineering, SPP University, Pune, India[3]

U.G. Student, Department of Computer Engineering, JSPM'S Imperial College of Engineering, SPP University, Pune, India[4]

Professor, Department of Computer Engineering, JSPM'S Imperial College of Engineering, SPP University, Pune, India[5]

ABSTRACT: This paper describes a natural language system which is connected to a database system. The natural language system accepts a user input in natural language via voice input. Then it extracts the necessary information needed for the formation of the SQL Query. The extraction or cleaning process is based on the set of keywords we defined. After which we give this information to a multinomial logistic regression classifier, which predicts which type of query is requested by the user. This information is further used to form the final query and output in given to the user on the interface. We feel such easy to use and high-level interfaces will be needed as information systems become more readily available to more users.

KEYWORDS: Natural Language Query, Speech-to-text, Speech Recognition, Logistic Regression, Structed Query Language (SQL), Database Query.

I. INTRODUCTION

The main purpose of energy efficient algorithm is to maximize the network lifetime. These algorithms are not just related to maximize the total energy consumption of the route but also to maximize the life time of each node in the network to increase the network lifetime. Energy efficient algorithms can be based on the two metrics: i) Minimizing total transmission energy ii) maximizing network lifetime. The first metric focuses on the total transmission energy used to send the packets from source to destination by selecting the large number of hops criteria. Second metric focuses on the residual batter energy level of entire network or individual battery energy of a node [1]. In this paper, we address the problem of automatic generation of Structured Query Language (SQL) queries. SQL is a database language for querying and manipulating relational databases.

Writing and executing SQL queries is an integral part of relational database courses. Natural Language Processing (NLP) is one of the most active techniques used in Human-Computer Interaction. It is a branch of Artificial Intelligence (AI) that is used for information retrieval, machine translation and linguistic analysis. The main objective of NLP is to allow communication between human and computers without memorizing commands and complex procedures.

Asking questions to databases in natural language is a very convenient and easy method of data access, especially for casual users who do not understand complicated database query languages such as SQL. This system focuses on the solution of the problems arising in the analysis or generation of Natural language text or speech.

Using our current system, we can predict which query the user has requested for, is it a SELECT, UPDATE, DELETE, or any other query for that matter. This prediction and training the model to give the correct prediction is what is the most import part. After this it will form the final SQL query based on its type and execute it.

The remainder of the paper is organized as follows. In section 2 we have a brief explanation of our literature survey. Section 3 give us an idea of any existing systems. Section 4 presents a brief description of the proposed system. Section

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 6, June 2020||

5 details the architecture of the system. Section 6 explains the algorithm is detail. Finally, section 7 and 8 presents the conclusions and references.

II. LITERATURE SURVEY

1.Title: A Model of a Generic Natural Language Interface for Querying Database

Author:BaisHanane and MustaphMachkour

Abstract: They made a model for Natural language processing using database (NLDBI). This model is based on machine learning for querying database which improves knowledge based on machine learning approach. They showed two approaches for this,

Linguist Component: Which performs three analysis morphological, syntactic and semantic.

Database Knowledge Component: Where it consists of two parts DBQ generation and DBQ execution. The task of the DBQ generation is to translate the IXLQ created by the semantic analyzer into SQL. By mapping each element of the logical query to its corresponding clause in the SQL query.

Once the DBQ is generated it will be executed by the Database Management System (DBMS), and then, displays the answers returned in tabular form

2.Title: Database Query Formation from NL using semantic Modelling and Statistical keyword Meaning Disambiguation

Author: Frank Meng and Wesley W. Chu

Abstract: Here, they show how a NLP interface which will allow users to supply query information from NL input. They used High-level Query Formulator to access the semantic graph which also composes a formal database query in the end.

N-gram vectors where used to capture the lexical content like converting the natural language sentence into tokens and measuring if they have any meaning in database language.

3.Title: A Natural Language Database Interface Based On A Probabilistic Context Free Grammar

Author: Bei-Bei Huang, Guigang Zhang, Phillip C-Y Sheu

Abstract: This paper presents a natural language interface to relational database. It introduces some classical NLDBI products and their applications and proposes the architecture of a new NLDBI system including its probabilistic context free grammar, the inside and outside probabilities which can be used to construct the parse tree, an algorithm to calculate the probabilities, and the usage of dependency structures and verb subcategorization in analyzing the parse tree. Some experiment results are given to conclude the paper.

4.Title: Natural Language Interface to Database Using Co-occurrence Matrix Technique

Author: Anuradha Mohite, VarunakshiBhojane

Abstract: This paper showed how data stored in database can be accessed by using SQL queries. Those who are expert in SQL language can access information from database but non-technical user cannot retrieve data from database such as MySQL. There was a need to provide natural language interface to database for non-technical users. In this paper they have discussed the design and implementation of a system using modified word co-occurrence matrix method which will provide access to database using queries in English language

5.Title: Automatic SQL Query Formation from Natural Language Query

Author: PrasunKanti Ghosh, Saparja Dey, Subhabrata Sengupta

Abstract: Here they explained the process carried out by the Natural Language Processing system by means of a method known as "Levels of Language" or Synchronous Model of language. It was divided into four stages such as Morphology, Lexical, Syntactic, Semantic. The stages had their own significance such as, breaking down the sentences into tokens, after which interpret the meaning of individual words in which all the tokenized sentences will be mapped with the meaning of the same word.

After which they found the attributes present in the input query from the words generated in the previous stages. Semantics focuses on the study of meaning of the words present in the natural language query and the relation between signifiers like words, signs, phrases and what do they actually stand for. And they used speech recognition using python for android as it was an android based project.

6.Title: A Rule Based Approach for NLP Based Query Processing.

Author: Tanzim Mahmud, K. M. Azharul Hasan, Mahtab Ahmed, ThwoiHla Ching Chak

Abstract: Databases and database technology are having major impact on the growing use of computers. In order to retrieve information from a database, one needs to formulate a query in such way that the computer will understand and

produce the desired output. But the non-IT people cannot be able to write SQL queries as they may not be aware of the SQL as well as structure of the database. So, there is a need for non-expert users to query the databases in their natural language instead of working with the values of the attributes. This paper gave an idea for accessing the database easily using natural language without having any knowledge about the query language. The approach is a rule-based approach. The obvious advantage is that it makes a great promise for computer interfaces easier for the use of general people. Because of this, people will be able to communicate to the computer in their own language instead of learning a specialized language or commands.

7.Title: Formation of SQL from Natural Language Query using NLP

Author: Uma M, Sneha V, Sneha G, Bhuvana J, Bharathi B.

Abstract: This paper gave an insight to how a system using NLP by giving structured natural language question as input and receiving SQL query as the output, to access the related information from the railways reservation database with ease. The steps involved in this process are tokenization, lemmatization, parts of speech tagging, parsing and mapping. They have achieved 98.89 per cent accuracy. This paper gives an overall view of the usage of Natural Language Processing (NLP) and use of regular expressions to map the query in English language to SQL.

8.Title: Review on Natural Language Processing and its Toolkits for Opinion Mining and Sentiment Analysis.

Author: Yasir Ali Solangi, Zulfiqar Ali Solangi, SamreenAarain, Amna Abro, Ghulam Ali Mallah, Asadullah Shah

Abstract: In this paper, Natural Language Processing (NLP) techniques for opinion mining and sentiment analysis are reviewed. Initially NLP is reviewed then briefed about its common and useful pre-processing steps also. In this paper opinion mining for various levels are analyzed and reviewed. At the end issues are identified and some recommendation are suggested for opinion mining and-sentiment-analysis.

III. EXISTING METHODOLOGY

1. There are existing systems such as ELIZA which is a simulation of a Rogerian psychotherapist.
2. SHRDLU (Terry Winograd, 1968) was an early natural language understanding computer program, developed by Terry Winograd at MIT in 1968–1970.
3. And LIFER/LADDER (Hendrix, 1978) was one of the first good database NLP systems. It was designed as a natural language interface to a database of information about US Navy ships.
4. Since then a lot of advancements have been made in NLP, like the use of NLDBI (Natural Language Database Interface) and the use of CFG (Context free grammar). These systems have been already implemented and we propose a new way to use NLP for querying database systems.

IV. PROPOSED SYSTEM

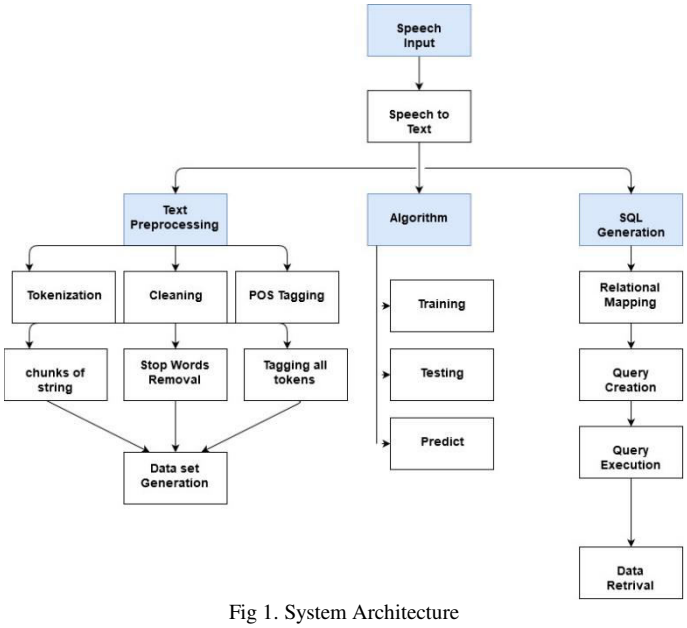

Fig 1. System Architecture

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 6, June 2020||

Nowadays data is increasing rapidly. There are so many new database tools and technologies are growing, therefore we can store large data, but the problem is that the technology or an interface which can process data and display the data as per the user request is not familiarized with many of the people.

It means many people don't have proper knowledge of handling database. So, we are implementing a system which will be useful to convert natural language questions into SQL query so that user can access exact data from database without prior knowledge of database.

So, there are multiple stages to how this current system or ours works, it includes:

1. **Voice Recognition:** User will give voice input, which will be recognized and then converted into text format.

2. **Text processing:** Perform pre-processing on the text converted from voice

- Tokenization i.e.; In this phase, the sentence is broken down into tokens. Here, we split the given input query sentence in natural language into all the words it contains and store the words in a list.
  - Example: Show me all the students from B.E
  - It will get converted into: [“Show”,” me”,” all”,” the”,” students”,” from”,” B.E”]
- Stop words Removal: Stop words like; I, me, we, here, you, etc. will be removed from the tokenized list.
  - The list in currently in this state, [“Show”, “me”,” all”,” the”,” students”,” from”,” B.E”]
  - After stop words removal the output will be, [“Show”,” students”,” B.E”]
- Parts of speech tagging: It can be important for syntactic and semantic analysis. So, for something like the sentence above the word can has several semantic meanings.

3. **Multinomial Logistic Regression Algorithm:** This will predict the type of query from the information that is given to it from the previous steps.

4. In the next step we will validate which query to execute, after which the query is generated and executed.

5. Data is fetched from the database and then it is displayed to the user on the interface.

IV. IMPLEMENTATION

System is provided with a main interface where user can register or log in system for access.If the admin is not Registered then using registration user create an account in system.Authorized Registered User Can Log in system to access the data. The user interacts with the system via Voice Commands by speaking his/her Natural Language Query for the further output. The query is based on casual human speech or conversation. The spoken query undergoes many steps to arrive at the final results. It includes synonym table which is used to convert the spoken query into SQL keywords.

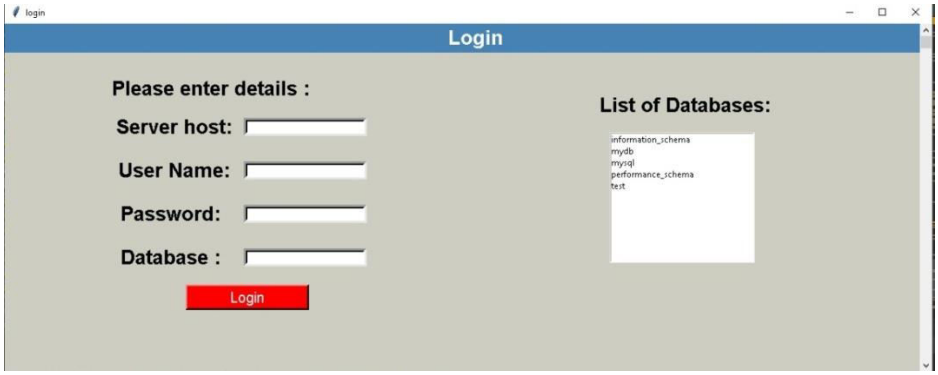

Fig.2.Login Page.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 6, June 2020||

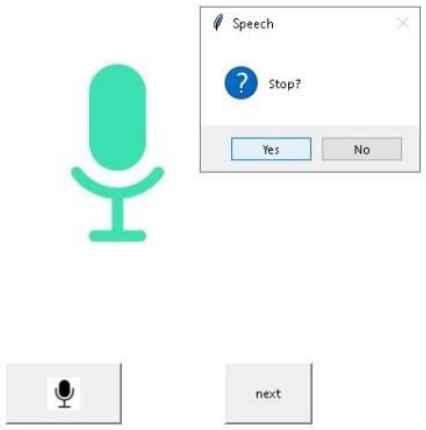

Fig. 2. Voice Input

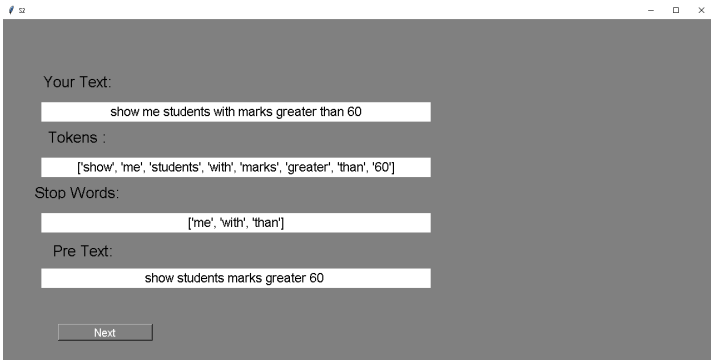

Fig. 3. Text Processing.

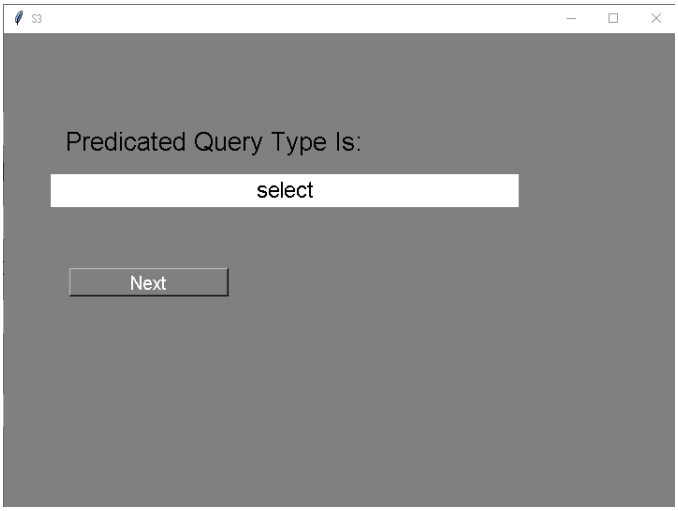

Fig 4. Predicted Query Type.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 6, June 2020||

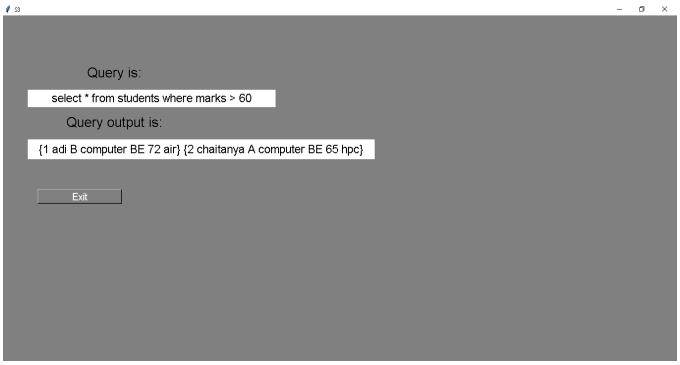

Fig 5. Predicted and Extracted Query.

IV. RESULT

In this proposed system, machine learning approach is used to predict type of query. Predicting type of query come under classification problem therefore multinomial Logistic Regression is used for predicting the type of query. On the Basis of tagged tokens, the noun map and verb list is prepared through one iteration over the tokens. then the tokens are given a unique identifier using Label encoder and the data is trained on logistic regression algorithm. then model will predict whether the natural language statement represents a data retrieval query (SELECT) or a DML query (ALTER, DELETE) is taken at this stage with the help of certain data arrays for denoting type of query. For example, when words like select and its certain synonyms appear in the input, the type of query is predicted as select as per the trained model and so on. The model predicts the query type accurately with accuracy of 98.65.

Query Type	No of Sample tested	Accurate Prediction
Select	34	34
Alter	17	17
Delete	12	12
Drop	4	4

Table 1. Result analysis based Trained model

The system has stored training Samples in Dataset which is a Data frame consisting of two columns i.e. words and Query type. as per the analysis and training, more the sample stored in Data frame, more query type gets predicted accurately and query gets successfully executed. As in Table, initially only 15 samples of data and query type where stored in Data frame, only five queries got executed. samples are increased to 20, then 8 queries got successfully executed. For executing 30 plus queries successfully, at least 45 and more samples must be stored in the Data frame. As the number of words and types stored in Data frame increases, the possibility of executing the query successfully gets increased.

No. of Samples Stored in Data frame	No of Query executed
15	5
20	8
25	10
30	15
35	21
40	28
45	34

Table 2. Result analysis based on Sample stored in Data frame

The confusion matrix is shown in following diagram.

Result Analysis on Logistic Regression Model Training and testing:

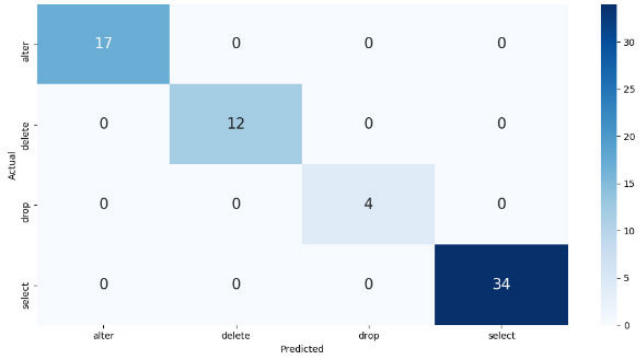

Fig 6.Confusion Martix

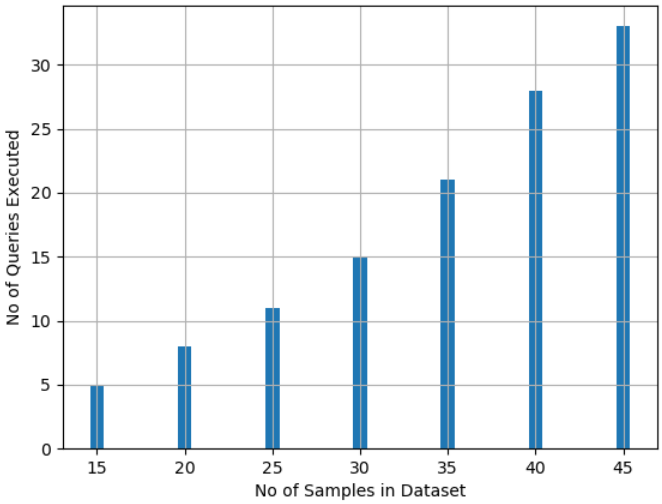

Fig 7.Result analysis based on Sample stored in Data frame

V. CONCLUSION

This could as well be a step forward to making database usability accessible to people who have no knowledge how a database query works. Natural Language Processing can bring powerful enhancements to virtually any computer program, because human language is so natural and easy to use for humans. Various processes like tokenization, syntactic and semantic analysis are carried out to generate an equivalent SQL query from a natural language query. This system predicts what the user query is and then validates the query, which is then executed. To get the maximum performance, the data dictionary of the system will have to be regularly updated with words that are specific to the particular system. This system is currently capable of handling simple queries along with some complex queries. Because not all forms of SQL queries are supported, further development would be required. Using our system any novice user can handle a database system efficiently and with ease.

REFERENCES

1. Uma, V. Sneha, G. Sneha, J. Bhuvana and B. Bharathi, "Formation of SQL from Natural Language Query using NLP," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, India, 2019, pp. 1-J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 6, June 2020||

2. Bais, Hanane, Mustapha Machkour and Lahcen Koutti. “A Model of a Generic Natural Language Interface for Querying Database.” international journal of intelligent systems and applications. 8. 35-44. 10.5815/ijisa.2016.02.05.
3. Frank Meng and Wesley W. Chu, “Database Query Formation from Natural Language using Semantic Modelling and Statistical Keyword Meaning Disambiguation”, 1999.
4. T. Mahmud, K. M. Azharul Hasan, M. Ahmed and ThwoiHla Ching Chak, "A rule based approach for NLP based query processing," 2015 2nd International Conference on Electrical Information and Communication Technologies (EICT), Khulna, 2015.
5. A. Mohite and V. Bhojane, "Natural language interface to database using modified co-occurrence matrix technique," 2015 International Conference on Pervasive Computing (ICPC), Pune, 2015, pp. 1-4.
6. Ghosh, Prasun&Saltlake, Kolkata & Kolkata, Saparja& Dey, Kolkata & Sengupta, Subhabrata& Assistant, Kolkata &Saltlake,. (2014). “Automatic SQL Query Formation from Natural Language Query”, International Conference on Microelectronics, Circuits and Systems (MICRO-2014).
7. Y. A. Solangi, Z. A. Solangi, S. Aarain, A. Abro, G. A. Mallah and A. Shah, "Review on Natural Language Processing (NLP) and Its Toolkits for Opinion Mining and Sentiment Analysis," 2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS), Bangkok, Thailand, 2018, pp. 1-4.
8. B. Huang, G. Zhang and P. C. Sheu, "A Natural Language Database Interface Based on a Probabilistic Context Free Grammar," IEEE International Workshop on Semantic Computing and Systems, Huangshan, 2008, pp. 155-162.
9. Devendra P Gadekar, Dr. Y P Singh,” Content Based Filtering and Fraud Detection on Social Networking Sites” Journal of Advances in Science and Technology Vol. 15, Issue No. 1, March-2018, ISSN 2230-9659.