Construction of SQL Query Using NLP Through Voice Recognition

Persis K. Chandane, Dept. of Computer Engineering,MES College of Engineering, Pune,India.

Nikita S. Sonavane, Dept. of Computer Engineering,MES College of Engineering, Pune,India.

Snehal R. Ingale, Dept. of Computer Engineering,MES College of Engineering, Pune,India.

Ashwini P. Kamble, Dept. of Computer Engineering,MES College of Engineering, Pune,India.

Shilpa P. Khedkar, Dept. of Computer Engineering,MES College of Engineering, Pune,India.

---------------------------------***---------------------------------

Abstract

The Natural language processing is a field of computer science anxious with the connections among computers and human (natural) languages like English and other languages. It is becoming one of the utmost active areasin the communication between human and computer. These include spoken language systems that combine speech and natural language.Extracting information from database is characteristically done by using a structured language such as SQL (Structured Query Language). But non expert userscan't use this later. Wherefore using Natural Language (NL) for example English sentences for communicating with database can be a powerful tool. It is an rapidly growing research area at the border between linguistics Data and artificial intelligence, aiming at developing computer programs capable of human-like activities like interpreting or producing texts or speech in a natural language, such as English or adaptation of natural language in text or speech form to languages similar SQL. The most important submissions of natural language processing include information retrieval and information organization, machine translation. The main aim of NLP is to enable communication between people and Machines lacking the resorting to memorization of complex commands and procedures.

Keywords: Natural language Query; Speech-to-text; log file; Data dictionary.

1.Introduction

The storage of data is a crucial task in today's commercial system especially social media, database size is increased and accessing data from database become more crucial part in the recent research world.There are 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 request is not familiar with many of the normal peoples. Natural language processing (NLP) is becoming most active techniques to process on human language. In case of social media data extraction, the query conversion is very crucial task in terms of receiving exact data which is demanded by the users. The query can be of simple English language statement,these statements must be converted into proper SQL statement so that exact data can be fetch from database. so, these factors are taken as an evidence for implementing the proposed system. The objective of NLP is to facilitate statement among human and computers without complicated instructions and procedures. In other words, NLP is the technique that can used the natural languages like English. An end user can be easily processing their query without knowledge of Structured Query Language. Therefore, in this system the communication with the database is implemented for simple English language and analyzed for the accuracy. This gives access to a user to input their queries in simple English and get the answer in same language which is referred as Natural Language Interface to a Database. Thus, we have done successful implementation of SQL query generation from the natural language statement.

2.Literature Survey

As today's world is moving from offline to online all the process now are been done with the help of computers, any data we require is present on the internet so it's a challenging task to develop a system that will provide search interface/NLP System for users without knowing any specific syntax or knowledge of a database language. Hence, we present a system that will provide the search interface for users especially for online applications, search engines and many other diverse databases, where accuracy and efficiency are most significant terms required. this System is not able to generate query if input words are in short forms.

In Semantic analysis of natural language queries using domain ontology for information access from database, This System describes a method for semantic analysis of natural language queries for Natural Language Interface to Database (NLIDB) using domain ontology. Implementation of NLIDB for serious applications like railway inquiry, airway inquiry, corporate or government call centers requires higher precision. This can be achieved by increasing role of language knowledge and domain knowledge at semantic level. Also design of semantic analyzer should be such that it can easily be ported for other domains as well. In this paper a design of semantic analyzer for railway inquiry domain is reported. Intermediate result of the system is evaluated for a corpus of natural language queries collected from casual users who were not involved in the system design.
System is not able to generate query if input words are in short forms.

The goal of NLP (Natural Language Processing) is to enable communication between people and computers without requiring to memorization of complex commands and procedures. In other words NLP (Natural Language Processing) is techniques which can make the computer understands the languages naturally used by humans. The main purpose of natural Language Query Processing is for an English sentence to be interpreted by the computer and appropriate action taken. 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. System is not able to generate query if input words are in short forms. This paper proposes a method of querying with the databases by means of a natural language interface. This is hot issue in the area of database management is to provide a high-level interface for nontechnical users. Normal users are not aware with the formal language like SQL.

Then the problem is how they interact with the database system. A normal user may find him/her self-handicapped to deal with the database system. The paper presents an interface module that converts users query given in natural language into a corresponding SQL command. . Asking questions to databases in natural language like English is a very convenient and easy method of data access from database system, especially for normal users who do not understand complicated database query languages such as SQL. This paper proposes the architecture for translating English Query into SQL. Keywords: Databases, Database Management System (DBMS), Structured Query Language (SQL), Natural Language Interface for Databases (NLIDB), Natural Language Processing (NLP) System is not able to gener ate query if input words are in short forms.

3.Proposed System

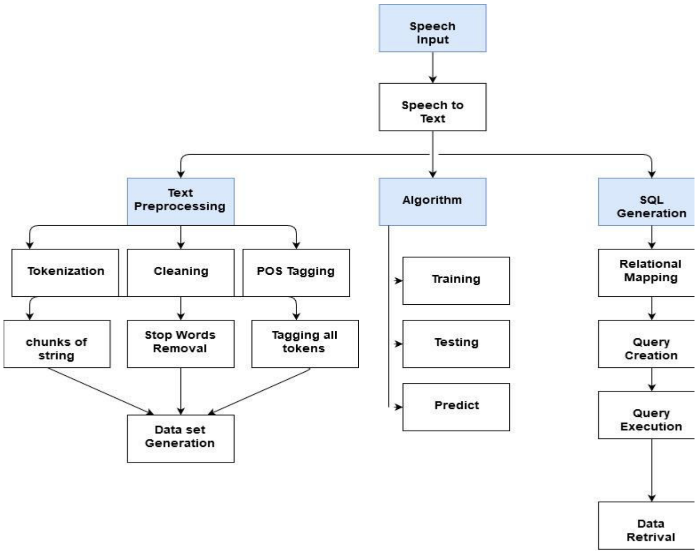


Fig 1: System Architecture

User will give voice commands to start system through microphone. Voice will get recognized by system. then it will get converted in text. Then input natural language query gets split into different tokens with the help of the tokenizer, word tokenizer, from NLTK package. The tokenized array of words is tagged according to the part-of-speech tagger using the Stanford POS tagger. The extra/stop words are removed which are not needed in the analysis of query. On the basis of the tagged tokens of earlier step, the noun map and verb list is organized through one iteration over the tokens. then the tokens are given a unique identifier using bag of words and thedata is trained on logistic regression algorithm.

then model will predict whether the natural language statement represents a data retrieval query like select i.e. “select * from your table name” or a DML query (INSERT Query, UPDATE Query, DELETE Query) is taken at this stage with the help of certain data arrays for representing type of query. For example, when words like insert query and its certain synonyms appear in the input, the type of query is INSERT and so on. In any type of query, the uncertain tags S (SELECT), W(WHERE), O (ORDER BY) are plotted to the nouns representing the clauses to which they belong.

Using the noun map and verb list, the table set is prepared, which will hold the tables that are required in the query to be formed. This is based on the fact that the table names are any nouns or verbs. The noun map is used to discover the attributes which are needed in the final query. The attributes, the table related with the attribute and the clause tag are kept in an attribute-table map which is used in the final stage of SQL query formation.

After the relations, attributes and clauses **are mined, the final query is built on the** data. The query is then executed from MYSQL database and final query result is displayed on the user on the Graphical User Interface this way the user can access the data.

4.Methodology

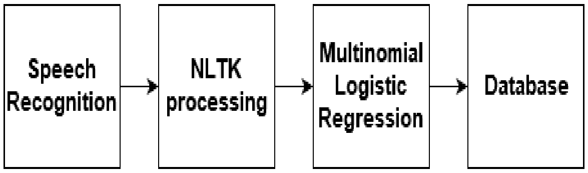


Fig 2: System block Diagram

1. Speech Recognition:

Speech Recognition is an important feature in numerous applications and technology is used such as home automation, artificial intelligence, and etc. This System uses the Speech Recognition library of Python. It is used to recognize the voice input.

2. NLTK

NLTK3 library for python will be used for input stemming, stop words removal and tagging. This library serves as a toolkit for computational linguistics. Token module delivers basic classes for processing individual elements of text, such as words, or sentences. NLTK tokenizer is used to tokenize incoming sentences and POS tagging is also done with NLTK.

3. Multinomial Logistic Regression:

Multinomial Logistic Regression is the regression analysis study to Perform when the dependent variable is nominal with more than two levels. This algorithm is used to predict the query type from processed text.

4. Database Connectivity:

MySQL is an open-source relational database management system. the records are stored inside the tables in MySQL database system. MySQL Server 5.5 is used and for python to MySQL connectivity PyMySQL Python Library is used.

5. 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.there are two options for taking input from user. user can give input using voice commands and text sentence as well. For text input a input prompt is provided. User can give input to the system using voice commands. 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.

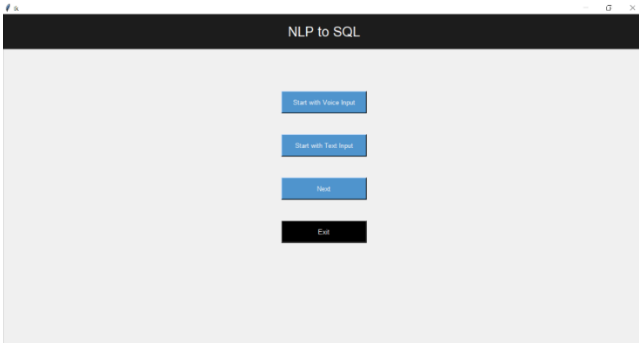
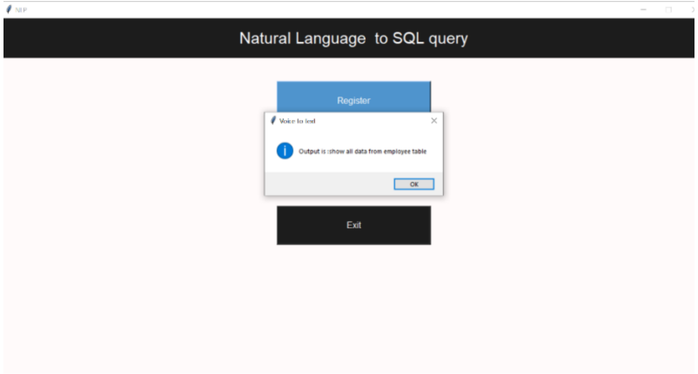


Fig. Input interface



Output is :show all data from employee table

Fig: Voice input

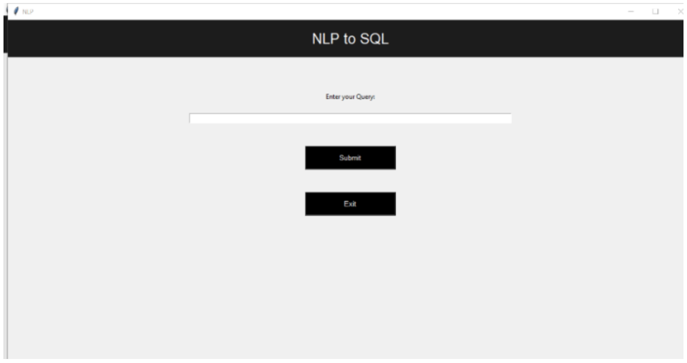
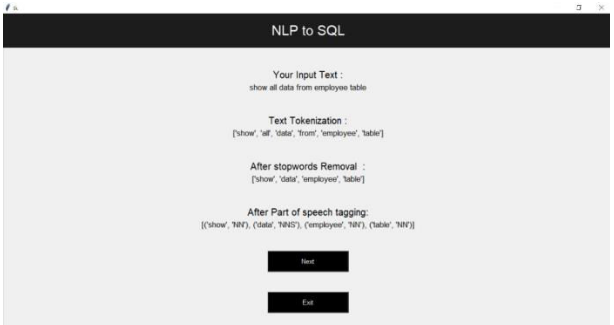


Fig: Text Input

Text Processing: Voice is converted to text and the text is tokenized. The query after lowercase conversion is thentransformed into stream of tokens and a token id is providing to each word of NLQ.The extra/stop words are removed which are not needed in the analysis of query. This is a veryimportant step which re formats the natural language query input string by removing the stop-wordslike I,me,am etc.The tokens are then classified into nouns, pronouns, verb and string/integer variables. In thissystem Part-Of Speech Tagger (POS Tagger) is a piece of software that reads text in some languageand assigns arts of speech to each word (and other token), such as noun, verb, adjective, etc., althoughgenerally computational applications use more fine-grained POS tags like noun-verb-plural.



Your Input Text :
show all data from employee table

Text Tokenization :
['show', 'all', 'data', 'from', 'employee', 'table']

After stopwords Removal :
['show', 'data', 'employee', 'table']

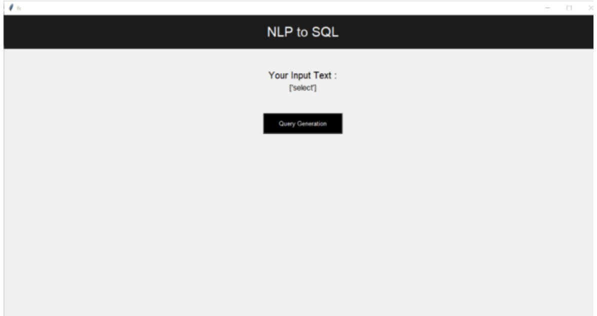
After Part of speech tagging:
[('show', 'NN'), ('data', 'NNS'), ('employee', 'NN'), ('table', 'NN')]

Fig: Text Processing

Query Type Prediction: Multinomial Logistic Regression is used for predicting type of SQL query. Based on the tagged tokens of earlier step, the noun map and verb list are prepared through one iteration over the tokens.

1.Two phases are there to build machine learning module testing and predict.
2. Identify the relation between Input and respective algorithm to predict the accurate results which
is stored in dictionary datasets and it's also corrects the formation of sentence.

3. After the relations, attributes and clauses are extracted, the final query is built.

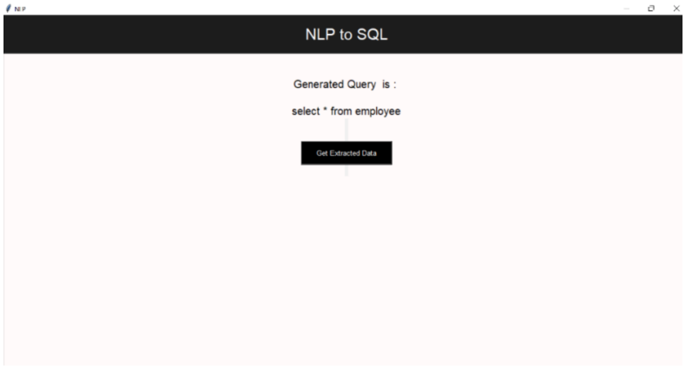


Your Input Text :
['select']

Fig: Predicted Query Type

Query Generation: After type prediction the Query is generated as per SQL syntax.
• After predicting type of query using multinomial logistic regression algorithm the

cleaned textis send to respective query generation method.In respective query generation method the query is constructed as per the SQL syntax.The Query is executed and the Data is Extracted. Data is fetched from database and displayed tothe user. The final query result is displayed to the user on the Graphical User Interface.

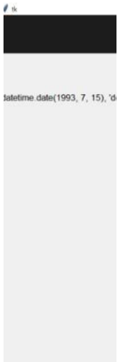


NLP to SQL

Generated Query is :

select * from employee

Fig: Generated Query



```
datetime.date(1993, 7, 15), 'd
```

Fig: Extracted Data

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

6.Result

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

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: 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 inTable, initially only 15 samples of data and query type where stored in Data frame, only five query gotexecuted. samples are increased to 20, then 8 queries got successfully executed. For executing 30 plusqueries successfully, at least 45 and more samples must be stored in the Data frame. As the numberof words and types stored in

Data frame increases, the possibility of executing the query successfullygets increased.

Table: 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:

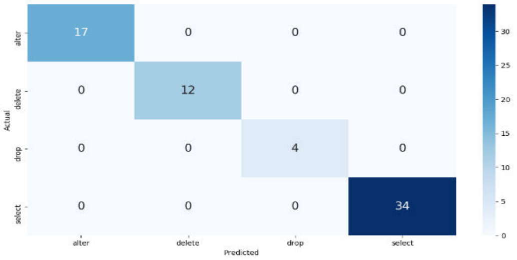


Figure: Confusion Matrix

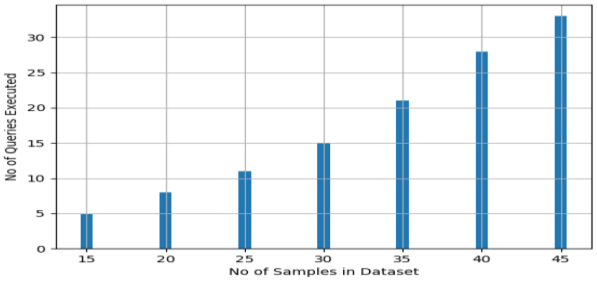


Figure: Result analysis based on Sample stored in Data frame

7. Conclusion

Repossessing information from the database requires knowledge of technical languages like SQL (Structured Query Language). In this project, we consider a NLP approach of interpreting English queries into equivalent SQL queries. In this approach we look at extracting certain keywords and indicators from an English query spoken, and then using a system to generate them query based on the keywords. And can perform different type of queries like select insert etc. At present our project is built around the fixed database and fixed language i.e. it can understand the queries about the fixed database only. But the project can be extended to cover any database. Support for other languages should be provided for enhancing the project. The program should identify the vendor of the selected database and should use appropriate drivers to connect them.

8.References

[1] SatavAkshay, G.Archana B. Ausekar, Radhika M Bihani and Mr AbidShaikh, 2014.A Proposed Natural Language Query Processing System, International Journal of Information Technology, 3(2).

[2] Avinash j. agrawal and o.g. kakde, 2013. semantic analysis of natural language queries using domain ontology for information access from database i.j.intelligent systems and applications.

[3] Gaikwad Mahesh, P., 2013. Natural Language Interface to Database, International Journal of Engineering and Innovative Technology (IJEIT),2(8).

language processing, international journal of computing and business research.

[5] Natural Language Web Interface for Database By Rukshan Alexander, Prashanthi Rukshan, and Sinnathamby Mahesan.

[6] Natural language to sql conversion System by Anil M. Bhadgale, Sanitha R. Gavas, Meghana N. Patiland Pinki R. Goyal PVGS COET, Pune, Maharashtra, India.

[7] Natural Langauge Query Processing Using Probabilistic Context free GrammerByArati K.

Deshpande1 and Prakash. R. Devale2 1Student and 2Professor and Head, Department of Information Technology, BharatiVidyapeeth Deemed University, Pune, India.

[8]Natural language to SQL Generation for Semantic knowledge Extraction in Social Web Sources By K. JavubarSathick and A.Jaya.

[9]Natural language Interface for Database: A Brief review Mrs. NeeluNihalani , Dr. Sanjay Silakari , Dr. Mahesh Motwani.

[10]An algorithm to transform natural language into SQL queries for relational databases Garima Singh, ArunSolanki.