NLP TO Create SQL Query

Prof.Khan Tabrez[1] Shaikh Shagufta[2] Shaikh Sharmeen[3] Momin Ummiya[4] Shaikh Rameeza[5]
[1,2,3,4,5]Department of Computer Engineering
[1,2,3,4,5]Anjuman-I-Islam Kalsekar Technical Campus School of Engineering

Abstract—NLP to create sql query, this intelligent system converts the human language into the structured query language. Every year, every minute thousands of data is generated and managed. To use this data or retrieving information, database interaction is important and for this purpose expertise are required. The problem is that it restricts the interaction between the naive user and the database. Only few people who have knowledge of formal database language can retrieve the desired information from the database. To overcome such a problem this proposed system would have the capability to analyze the user statements written in different ways and accordingly it gives response to the user. In this way it help a normal educated person having no knowledge of query language to easily interact with the system.

Key words: NLP, Morphoms

I. INTRODUCTION

In the present computing world, computer based information technologies have been extensively used to help many, private companies, academic and education institutions to manage their processes and information systems. Information systems are used to manage data. The information management system that is capable of managing several kinds of data, stored in the database systems. is known as Database Management System (DBMS). Databases are comprehensive element in private and public information systems which are essential in number of application areas. Databases are built with the objective of facilitating the activities of data management in information systems. Due to the progress and in deep applications of computer querying system. It is due to the fact that the technology in several areas to be accurate, databases have become the repositories of huge volumes of data .In relational databases, 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.

II. KEYWORDS

NLP (Natural Language Processing, IQR (Intermediate Query Representation), Morphoms (individual word, token (broken words), SQL (Structured Query Language).

III. PROBLEM AND SOLUTION

The limitation of such programs is that it restricts the interaction between user and database to predefined set of queries. Only few people who have knowledge of database structure and formal database language (such as Structured Query Language (SQL) can retrieve the desired information from database. A novice user having no knowledge of database structure and formal database query language cannot retrieve desired information if it is not supported by well thought application. Hence, it is a need to improve human computer interface that allows people to interact with the database in their natural language (such as English).

IV. OBJECTIVE AND SCOPE

The objective of the proposed system is the interaction of the naive user and intelligent system. The proposed system would have the capability to understand the natural language, learn the meaning of the query and process it and give the desired result. Since normal educated user can't access the database to overcome this problem there is a need to translate the normal user language such as English into SQL query in order to get the result from the database, here user need not to learn anything related to database and can interact directly to the database in his/her known language.

Proposed system can be used in any application. Where there is need to store the data and retrieving of data is done by non-technical person or naive user, this system is useful:

School, Colleges, Hospital, Transport, Bank ,Government Office, Service Sector, Navy, Manufacturing database, Sensus system, Agriculture**,** Chat system, Business organization, Chemical Industies.

V. LITERATURE REVIEW

An Overview of NLIDB Approaches and Implementation for Airline Reservation System: [1]

This system was developed for Flight Reservation. A combination of Syntax Analysis and Intermediate Query approach is used for this system. Syntax Analysis performs syntactic processing and breaks the input sentence into its constituent parts and identifies the relations between the concepts. Intermediate Query approach allows to easily perform the mapping of concepts to an intermediate representation. The intermediate representation can be used even in case of database portability i.e. even if database is ported to another database.

- Weaknesses
- This system is not developed for DDL (Data Definition Language).
- It is not implemented for complex queries like nested, join, group, order, queries, etc.
- Modification and Deletion to the database in this system is not done by natural language.
- It is restricted to single user only.
- It is also domain dependent i.e limited for airline reservation system.

A. Solution

1) Proposed system is designed for DDL as well as DML statement.
2) It would have the ability to deal with complex queries.
3) It would deal with the ambiguity in the query.
4) It would be accessible by multiple users.
5) It would be domain independent i.e. applicable to any database applications.
6) There will se security for admin and normal user so

that data can be more secured.

VI. PROPOSED SYSTEM ARCHITECTURE

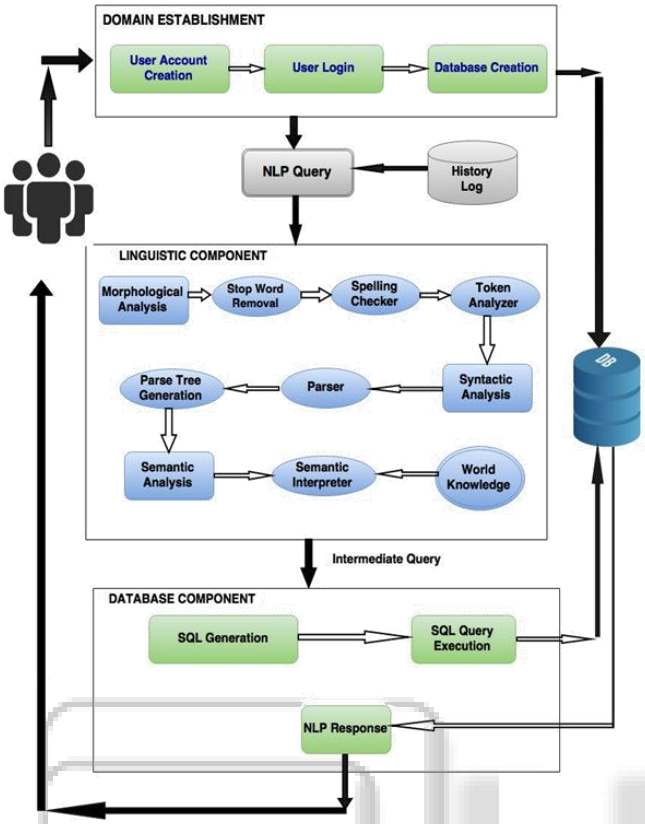

Fig. 1: System Architecture

The System Architecture consists of three modules:

- DOMAIN ESTABLISHMENT
- LINGUISTIC COMPONENTS
- DATABASE COMPONENTS

A. *Domain Establishment:*

This module is responsible for creating user accounts and database creation as the proposed system is would be used by any database application and by multiple users.

B. *Linguistic Component:*

This module is responsible for translating natural language input into a logical query. In this, the sentence is syntactically and semantically analyzed and processed, and an intermediate query is generated by the following steps.

- Morphological Analysis
- Syntactic Analysis
- Semantic Analysis

1) Morphological Analysis

Morphology in linguistics is the study and description of how words are formed in natural language. In this phase the sentence is broken down into tokens- smallest unit of words, and determine the basic structure of the word.[6]

For instance, unusually can be thought of as composed of a prefix un-, a stem usual, and an affix -ly. Composed is compose plus the inflectional affix -ed: a spelling rule means we end up with composed rather than composed.

2) Stop Word Removal:

Stop words are non-context bearing words, also known as noisy words which are to be excluded from the input sentence to speed up the process. For example again, already, amongst, etc. [6]

3) Spelling check:

Three methods for spelling check are as follows: [6]

Correct Token	Error Token	Spelling Checking Operation
student	dudent	Substitution
student	studnt	Insertion
student	studeent	Deletion

4) Token Analyzer: [6]

Each identified tokens can be represented as attribute token, value token, core token, multi-token, continuous token, etc.

- Attribute token- using metadata
- Core Token-first, all capital letters
- Numeric Token-digits , digits separated by decimal point
- Sentence Ending Markers- (. ? !)
- Value Token- (M.C.A, “mca”, ‘mca’)
- Continuous Token – (‘@’, apostrophe (‘) , ‘$’)
- Multi-token- emp_no or e-no
- Abbreviated Token- CE for Computer Engineering

5) Syntactic Analysis:

The objective of the syntactic analysis is to find the syntactic structure of the sentence.It is also called Hierarchical analysis/Parsing, used to recognize a sentence, to allocate token groups into grammatical phrases and to assign a syntactic structure to it.[2]

6) Parse Tree: [5]

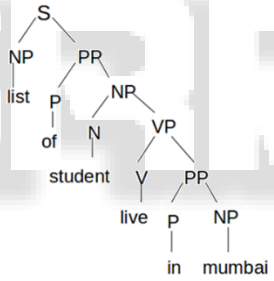

Fig. 2: Parse Tree

Parser generates a parse tree with the help of syntactic analysis. A parse tree or parsing tree is an tree in the ordered and structured form that represents the syntactic structure of a string according to some context free grammar.

Example: list of student live in Mumbai

7) Semantic Analysis:

Semantic Analysis is related to represent the meaning of linguistics sentences. It concerns with how to determine and understand the meaning of the each word. So, it is responsible for creating the logical query which acts as the input query to Database Query Generator. Hence this is another form of presenting the user tokens in the form of semantic word.[2]

8) IQR generation:

It becomes very difficult to map the syntax tree and semantic tree of the sentence directly to the sql query, an intermediate query is generated from the semantic analysis then this logical query is going to be converted into sql query. The logical query expresses the meaning of the user’s question in terms of user world knowledges, which are independent of the structure of the database. The logical query is then translated to database’s query language expression, and generates the user understandable result. [4]

C. Database Component:

1) SQL query generation:

It consists of SQL query generation where intermediate query is going to map to the sql query.

2) SQL query execution:

Generated sql query is going to be executed here and desired data is extracted from the database.

3) Response of NLP:

Extracted result is going to be displayed to the user.

VII. METHODOLOGY

Proposed system is based on Intermediate Representation Technique which is a combination of syntactic based system and semantic based system. System follows the following steps:

A. Domain Creation:

- User create the account
- User login to the system
- User creates the respective database
- User request the data from the database in English language

B. Query Analysis:

- This phase consists of morphological syntactic, semantic analysis of the sentence
- English query is split up into words.
- Morphoms are recognized from the sentence.
- Stop words are removed from the sentence.
- Error detection and correction of words will be processed
- Parser is responsible for the parse tree generation using NLTK library of Python.
- Part of speech tagging, sementic meaning of senetence takes place using Pattern library of Python.

C. Intermediate Query Generation [4]

Intermediate query is generated from above steps.For example: the question "What is the capital of each country bordering UP?" would be mapped to the following logical query.

answer([Capital, Country]):-
iscountry(Country),
border(Country, UP),
capitalof(Capital, Country).

D. Query Mapping: [4]

Mapping of words to sql attributes is done using Database Dictionary and semantic rules. Example of mapping and sql query generation for the above intermediate query. The mapping information could link the predicate is country to the Sql query. Example of above IQ. iscountry(Country),

SQL >> SELECT country FROM countries_table;

E. Response Generation:

- SQL query is generated and executed.
- Result is displayed to the user.

Example of the above query.

>>countries table;

Country
India
USA
Africa
Australia
Canada

VIII. COMPETITIVE ADVANTAGE OF PROJECT

There are many softwares are developed for processing the natural language to generate sql query and to extract data from database which are as follows:

A. Lunar:

It was involved in a system that answered questions about rock samples brought back from the moon.[5]

B. Liffer/Ladder:

It was designed interface of natural language to a database of information about US Navy ships.[5]

C. Chat-80:

It is designed to know the location of ocean, rivers, cities and countries. Semantic grammar technique is used in this system.[5]

D. Precise:

It is developed for Air Travel Information System & GEOQUERY. Lexical analysis and semantic constrains approach is used.

All the above systems are domain dependent system that means used for particular database application. These are based on shallow parsing, semantic grammar-based system, syntactic based system and it is very difficult to convert the natural language sentence to directly to the database query and the proposed system is domain independent.

Proposed system is based on IQR Technique which is internal logical query so it becomes easy to convert the logical query to the sql query. Another benefit is that existing systems does not maintain the history log of the query abd proposed system does. Besides these another benefit of the proposed system over existing system is that it can be used by many users.

IX. CONCLUSION

NLP for sql generation is very crucial aspect for naive and non-technical person to interact with the database system and this proposed system fulfills the requirements of the user to handle the database system.

System is designed that translate the english language to the sql in order to retrieve the data from the database. Propoesd system is domain independent that is it can be used any database application not restricted to particular application. Complex database queries can be evalutaed which are asked in natural language. Queries include order queries, join queries, nested queries, range queries, comparison predicates, conjunctions, quantifications, multi-level aggregations etc. It is designed for DDL and DML statements as well. System is based on IRQ technique which is internal representation query to and it is the combination of syntactic and semantic based system. System is also designed to deal with query logs in to reduce the duplicate interaction to the system to help the user to interact with the system.

X. FUTURE SCOPE

- To accept queries in vernacular languages.
- To include question based on prediction in case of Information Retrieval system. For example, user can ask question like: “when the student puja will complete the final year of her studies?”, “what will happen if a student fail”, etc.
- To support multimedia data such as image, sound and graphics can be attempted.
- To include computational phonology and text-to-speech.

ACKNOWLEDGMENT

We would like to extend our deepest gratitude to our HOD Prof.Tabrez Khan for extending his valuable guidance in the research. His mentoring gave way a direction to our work.

REFERENCES

[1] Manju Mony ,Jyothi M. Rao ,Manish M. Potey ,An Overview of NLIDB Approaches and Implementation for Airline Reservation System ,International Journal of Computer Applications (0975 – 8887) Volume 107 – No 5, December 2014
[2] F.Siasar djahantighi1, M.Norouzifard1, S.H.Davarpanah, M.H.Shenassa,Using Natural Language Processing in Order to Create SQL Queries ,Proceedings of the International Conference on Computer and Communication Engineering 2008 May 13-15, 2008 Kuala Lumpur, Malaysia .
[3] Dr. Paresh Virparia ,Amisha Shingala ,Design and Development of Natural Language Query Interface for Relational Databases.
[4] I. Androutsopoulos G.D. Ritchie P. Thanisch ,Natural Language Interfaces to Databases – An Introduction ,University of Edinburgh 80 South Bridge, Edinburgh EH1 1HN, Scotland, U.K.
[5] Pooja A.Dhomne1, Sheetal R.Gajbhiye, Tejaswini S.Warambhe Vaishali B.Bhagat ACCESSING DATABASE USING NLP ,IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308
[6] Ms. Amisha H. Shingala and Dr. Paresh V. Virparia, Research paper on Intelligent Natural Language Processor, (under press) in National Journal of Systems and Information Technology (NJSIT), ISSN:0974-3308.

