Generic Framework for Semantic Query Conversion in Social Websources

K.Javubar Sathick[1], Dr. A.Jaya[2], Mohamed Nayeemudeen[3]

[#] *Department of Computer Applications,B.S.Abdur Rahman University*

Abstract- **Many organizations and social networks use databases for storing information and they stored in the database in the form of tables. Data can be retrieved or accessed by SQL queries. End users of the system may not understand complicated SQL queries. This Research paper aims to derive an automatic query translator for Natural Language based questions into their associated SQL queries. This framework provides an user friendly interface between end user and the database for easy access of social web data from different websources such as facebook,twitter and linkedIn etc.,.This paper is implemented using java as the front end and SQL server as the back end. The main objective of this research work is to provide an optimized SQL query for the Natural Language question provided by the user.**

Keywords: Database Management System (DBMS), Structured Query Language (SQL), Natural Language Interface for Databases (NLIDB), Natural Language Processing (NLP),Social webdata.

I. INTRODUCTION

Database applications play an important role in today's commercial system especially with evolution of social media, the size of the database and data accessing pace become more crucial part in the recent research world.. Most of the businesses and social sites need these types of applications by using the SQL language. Natural language processing (NLP) is becoming one of the most active techniques used in Human-computer Interaction which plays a vital role since the social media started playing its part to a large extent in the current trend. It is a branch of AI which is used for Information Retrieval, Machine Translation and Language Analysis. In the context of social media the query conversion is quite important in terms of bringing out the exact data which is requested by the websuers who surf on the net. The query /request will be of natural statement such as blog,comment,tweets etc., these statement must be converted into a most reliable and acceptable form of typical query which system can understand and crack the exact data from the database. so these factors are acting as a precious evidence for the proposed framework through this article.The main goal of NLP is to enable communication between human and computers without memorization of complex Commands and procedures. In other words, NLP is the techniques that can make the computer to understand the natural languages used by humans. Today's requirement of commercial system is to extracting data from a DataBase Management System such as MS Access, Oracle and others. A person without knowledge of SQL may find himself/herself handicapped in corresponding with the database. Therefore in this work the development of system for people to interact with the database in simple English language is implemented. This enables a user to input their queries in simple English and get the answer in same language. This is known as a Natural Language Interface to a Database (NLIDB).

II. RELATED WORKS

The very first attempts at NLP database interfaces are just as old as any other NLP research. In fact database NLP may be one of the most important successes in NLP since it began. 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. The success in this area is partly because of the real-world benefits that can come from database NLP systems, and partly because NLP works very well in a single-database domain. Databases usually provide small enough domains that ambiguity problems in natural language can be resolved successfully. Here are some examples of database NLP systems: LUNAR (Woods, 1973) involved a system that answered questions about rock samples brought back from the moon. Two databases were used, the chemical analyses and the literature references. The program used an Augmented Transition Network (ATN) parser and Woods' Procedural Semantics. The system was informally demonstrated at the Second Annual Lunar Science Conference in 1971. LIFER/LADDER 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. This system, as described in a paper by Hendrix (1978), used a semantic grammar to parse questions and query a distributed database. The LIFER/LADDER system could only support simple one-table queries or multiple table queries with easy join conditions.

N. D. Karande, and G. A. Patil describes about,"Natural Language Database Interface for Selection of Data Using Grammar and Parsing",the authors proposed the NLDBI system considers selection of data and performing primitive queries onto database and join operation with some constraints. ATN (Augmented Transition Network) parser used for generating a parse tree.

Arati K. Deshpande and Prakash. R. Devale, proposed" Natural Language Processing using probabilistic

context free grammer",the author discussed a method to create new NLDBI system using probabilistic context free Grammar(PCFG). This paper highlights, Natural language statement is converted into internal representation based on the syntactic and semantic knowledge of the natural language. This representation is then converted into queries using a representation converter.

Dshish Tamrakar Dshish Tamrakar, published a article titled"Query Optimisation using Natural Language Processing", the author proposed the architecture for translating English Query into SQL using semantic Grammar. LIFER/LADDER method used in the syntax analysis. The LIFER/LADDER system could only support simple one tables Queries or multiple table queries with easy join conditions.

Alessandra Giordani and Alessandro Moschitti, proposed"Semantic Mapping Between Natural Language Questions and SQL Queries via Syntactic Pairing",the author proposed an automatic translation of natural language query into SQL query using support vector machine algorithm and kernel functions. In this algorithm to design a dataset of relational pairs containing syntax trees of questions and queries and encode them using kernel functions.

Gauri Rao, Snehal chaudhry, Nikita KulKarni, Dr.s.H.Patil, proposed" Natural language processing using semantic grammar", the author proposed the architecture for translating Natural language Query into SQL using semantic Grammar. Lexicon and post preprocessor used in the semantic analysis. Lexicon that stores all possible words that the grammar is aware of. Post preprocessor transforms the semantic representations of the sentence into a SQL query. This system capable of handling simple queries with standard joins conditions.

III. A SIMPLE ARCHITECTURAL LAYOUT

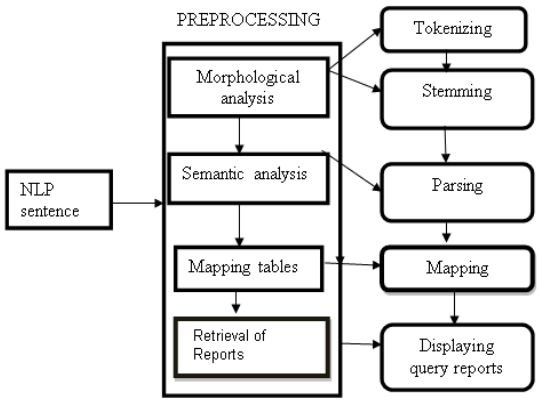

Fig 1.0 A Simple Architectural Layout

Fig 1.0 illustrates the simple architectural layout of the proposed framework.In this layout the Preprocessing phase consists of four modules such as Morphological analysis, Semantic analysis, Mapping table and Retrieval of reports. Here the Natural Language sentence is given as an input by the user. Morphological analysis involves the following steps.

(i)Tokenizing
(ii)Stemming

The second Semantic analysis performs parsing on the tokenized words. The third Mapping table maps the extracted words from the semantic analysis into their associated SQL queries. The fourth Generating SQL query generates SQL query and display the query reports.

IV. IMPLEMENTATION

The proposed framework is implemented by elaborating the actual working aspect of natural language processing analysis which is explained in few steps.Generally NLP has following steps:-

A. Morphological Analysis:

Individual words are analyzed into their components and non word tokens such as punctuation are separated from the words.

B. Syntactic Analysis:

Linear sequences of words are transformed into structures that show how the words relate to each other. Some word sequences may be rejected if they violate the languages rules for how words may be combined.

For example An English semantic analyzer would reject the sentence "Boy the go the to store ".

C. Semantic Analysis:

The structures created by the syntactic analyzer are assigned meanings. In other word mapping is made between syntactic structure and objects in the task domain. Structures for which no such mapping is possible may be rejected.

For example In most universes the sentence "Colorless green ideas sleep furiously" would be rejected as semantically anomolous.

D. Discourse integration:

The meaning of an individual sentence may depend on the sentences that precede it and may influence the meanings of the sentences that follow it.

For example The word "it" in the sentence "John wanted it" depends on the prior discourse context, while the word "John" may influence the meaning of later sentence (such as "he always had").

E Pragmatic Analysis:

The structure representing what was said is reinterpreted to determine what was actually meant.

For example The sentence “Do you know what time it is ?” should be interpreted as request to be told the time. The boundaries between these five phases are often very fuzzy. The phases are sometimes performed in sequence; One may need to appeal for assistance to another. For example part of process of performing the syntactic analysis of the sentence “Is the glass jar peanut butter?”is deciding how to form two noun phrases out of four nouns at the end of the sentence.

We consider a database SQL Server 2005. We have placed 3 tables in this SQL Server database. Novice users can not access contents of databases as they don‟t have knowledge of SQL language. That‟s why we proposed system which will enable user to access contents of databases using simple English language. Suppose we want comment of a facebook whose likes “Flipkart” then we have to form a SQL query: Select comments from facebook where flipkart=”like”; For a novice user it is not possible to form SQL query so using our system he / she can simply asks a question like “**What is name and comments of facebook who likes flipkart?”** In our daily life we always use a WH question that‟s why proposed system easily interprets WH questions and generates its relevant intermediate query. In addition with WH questions our system works with In Which, Total Number Of, On Which type questions. Fig 2.0 represents the structure of the system in which the NLP sentence gets processed at the respective levels and refined as a typical sql query. The refined sql query will fetch the desired data from the database which finally mapped to the user.

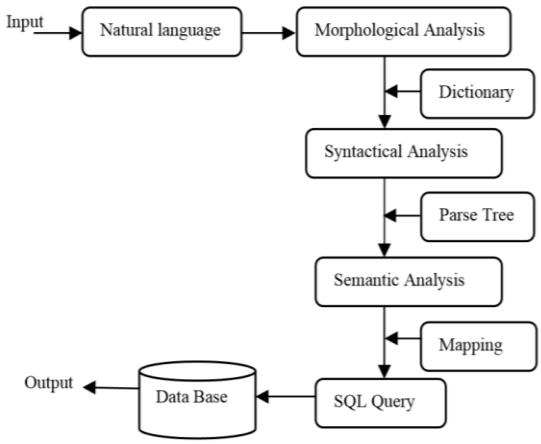

Fig.2 - Structure of the System.

Then the Morphological analysis are identified each world like what, is, the, Customer’s, FacebookId. Individual words are analyzed into their components, and it separated noun and adjective in the sentences. Morphological analysis must pull apart the word “Customer’s” into the proper noun “Customer” and the possessive suffix “’s”. A limited data dictionary is also used to store all related words about the system. After this, Syntactic rules checks the grammatical mistakes of a sentence and Semantic analysis must map individual words into appropriate objects in the knowledge base or database and the meanings of the individual words combine with each other and find out the meaning of simple English query. Example: Meaning of query: FacebookId of Customer with name Customer. Then the translator will change the above sentence with SQL query and with the help of SQL query, we will able to find out the results. SQL Query: Select custname, comments from facebook where jabong=”like”;

When user opens system he/she has to establish connection to database and then he/she can fire queries to database. User can asks queries to database in „How many‟, „Total number of‟, „ In which‟ format in addition with WH formats. Our system also provides facility to update tables in database. User can insert values into tables and can also delete values from table. Our system generates number of intermediate queries depending on semantics of user entered English statement. User have to select one of intermediate query which is more relevant to user‟s intended query. Then system will generate its appropriate SQL query. Our system also works fine with JOIN. User can retrieve data from two or more columns also.

Firstly system accepts English statement from user then system tokenized that statement and removes unwanted words. After that it identifies synonyms of column names and table names then replace synonyms with actual names. System places tokens in 4 parts depending on criteria words and then properly placing that parts generates one or more intermediate statements. This is one part of the system which only generates intermediate query. System simplify decision making task by relying on user for selection of intermediate query. This also helps to system to give proper output to the user and user can also easily recover from mistakes. After user selects intermediate query system‟s GenerateSQL module takes it as input and finds out 3 main keywords i.e. Select keyword, From keyword, Where keyword. Select keyword contains attributes that user wants to retrieve. From keyword contains table name from which user wants to retrieve attributes. From keyword can also contain more than one table then system has to generate query using JOIN as there is relationship between tables. Where keyword contains criteria which helps to retrieve specific contents by placing condition. Then GenerateSQL formats all these keywords in specific format and using different conditions that means formatting From keyword is different when there is only one table and different in case of two or more tables where we have to use JOIN. Then it places these keywords in standard SQL query and generate SQL.

"What is name and comments of facebook who likes flipkart" is processed by the system as given below.

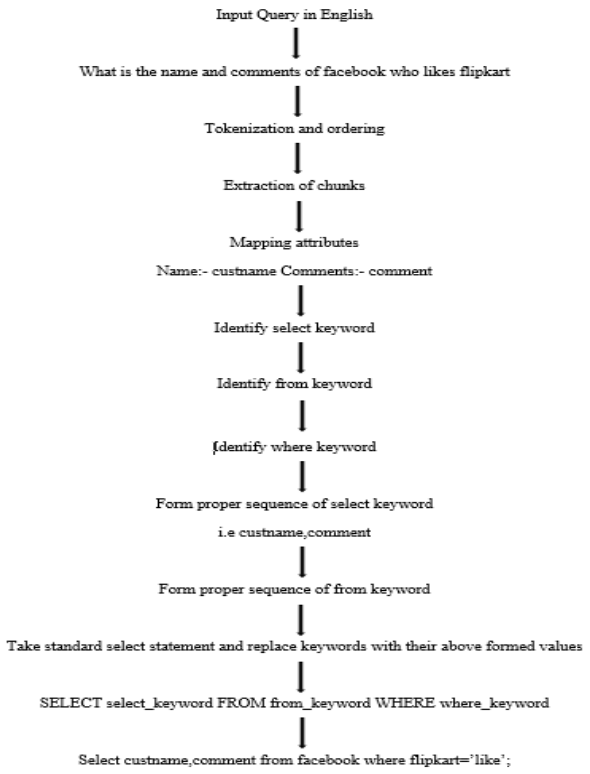

Fig 3.0 workflow diagram

Fig 3.0 illustrates the typical workflow diagram which explains the actual flow of process which is carried out to bring the well refined and appropriate query for fetching from the database.Now we consider some examples and see how system handles them. Assume our database contains two tables Twitter and Facebook in normalized form. Firstly we will take following example:-

What is the comments from Twitter who likes jabong?

Then by processing above English query system generates intermediate query i.e.

What is comment from twitter who like jabong

Generate SQL modules takes above query as an input and firstly finds out all attributes and table names then by interpreting meaning identify relation between tables and form query using JOIN condition. Output of GenerateSQL module for above query is as follows :-

Select count(comment) from facebook JOIN twitter fbid=tid where flipkart="like";

Analytical sequence of steps followed in the proposed framework:

- **Process Query**
  - Divide Query in tokens.
  - Remove punctuation marks.
  - Do initializations.
- **Generate Intermediate Analysis for query formation**
  - Divide Query into parts using criteria words.
  - Identify column attributes and table names from user Query and remove unwanted words.
  - Replace synonyms of column attributes and table names in Query with its actual names.
  - Arrange parts in proper sequence.
- **Formation of SQL Query**
  - Take intermediate Query as input
  - Identify/ derive 3 things from Query

- **Select keyword**: - These are attributes which user wants to retrieve.
- **From keyword**: - This is the table name from which user want to retrieve data.
- **Where keyword**: - This is condition specified in query.

- Replace select keyword with actual table attributes.
- If there is only one from keyword then Replace it with actual table name. else form following sequence table name1 JOIN table name2 ON attribute1(primary key of table1) = attribute2(attribute in table2 which is foreign key of table1)
- Replace where keywords attribute with actual table attribute and concatenate „=" following with value specified by user.
- Form standard template of SELECT Query and substitute above keywords i.e. select keyword, from keyword, and where keyword in their appropriate place.

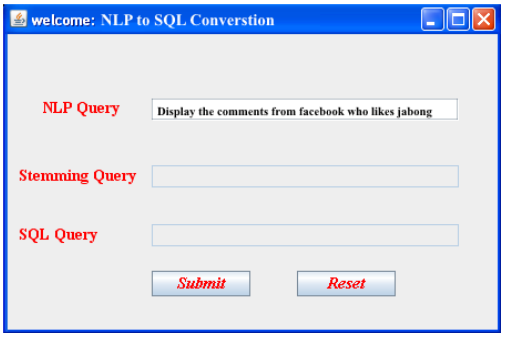

Figure 4.0 ScreenShot For translator

Figure 4.0 describes the design of the translator. The translator will accept NL sentence as input . when submit button pressed, it performs the translation of Natural language sentence into SQL . The reset button resets the window for next query.

Figure 5.0 ScreenShot For Stemming Process

Figure 5.0 describes the stemming process. This is done by using Porter algorithm. Stemming process used to identify the root word and remove the unwanted words in the given input.It will give the important keyword of the NLP sentence.

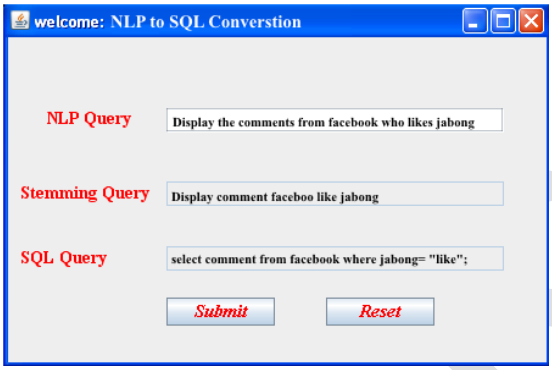

Figure 6.0 Screenshot for Displaying SQL query

Figure 6.0 displays the suitable SQL query for the Natural language sentence. It also displays the important keywords in the given input NLP sentence.

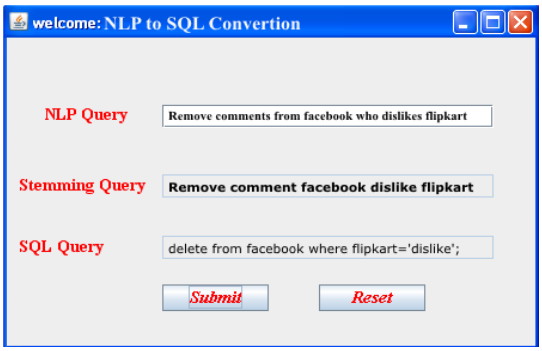

Figure 7.0 Screenshot for Removing records from the Table

Figure 7.0 illustrates the second type of input .here the user is giving NLP query for removing records from the table . When user clicks the submit button Then it will delete the appropriate records from the table and it will display the entire details except the removing details.

V. CONCLUSION AND FUTURE ENHANCEMENT

In this work a system is developed that is able to execute both DDL and DML queries, input by the user in his/her natural language (English). The system is developed in JAVA.Programming language and various tools of java are used to build the system. An oracle database is used to store the information's.Input given by the user is not required in the form of questions (who-form like what, who, where, etc). A limited Data Dictionary is used where all possible words related to a particular system are included. The Data Dictionary of the system must be regularly updated with words that are specific to the particular system. Ambiguity among the words will be taken care of while processing the natural language. The results show that our software is correct andhandlesthe SQL Queries without any problem.Future researches should consider factors that lead users to reformulate their NLP sentence. Also new research should be done to gather more information in various levels of understanding, effectiveness and situations. Method of gathering information from multiple tables should be carrying forward in the future.

REFERENCES

[1]. Anil M. Bhadgale, Sanhita R. Gavas, Meghana M. Patil & Pinki R, (Jun 2013), **Natural Language To Sql Conversion System,** Goyal PVG"s Coet, Pune, Maharashtra, India.

[2]. Saravjeet Kaur, Rashmeet Singh Bali , (2012) **Sql Generation and Execution from Natural Language Processing,** MMU, Mullana.

[3]. Arati K. Deshpande1 and Prakash. R., (May 2012), **Natural Language Processing using probabilistic context free grammer**, International Journal of Advances in Engineering & Technology, Devale ,Department of Information Technology, Bharati Vidyapeeth Deemed University, Pune, India,

[4]. Dshish Tamrakar, Deepty Dubey,(Mar-2012) **Query Optimisation using Natural Language Processing**, Dept. of CSE, Chhatrapati Sivaji Institute of Technology, CG, India.

[5]. Michael Gage, (Dec - 2012), **A Survey of Natural Language Processing Techniques for the Simplification of User Interaction with Relational Database Management Systems ,** California Polytechnic State University, San Luis Obispo

[6]. Alessandra Giordani and Alessandro Moschitti, (2010) **Semantic Mapping Between Natural Language Questions and SQL Queries via Syntactic Pairing**, Department of Computer Science and EngineeringUniversity of Trento Via Sommarive 14, 38100 POVO (TN) – Italy.

[7]. Gauri Rao Snehal chaudhry, Nikita KulKarni, (2010), **Natural language processing using semantic grammer**, Dr.s.H.Patil, (IJCSE) International Journal on Computer Science and Engineering.

[8]. N. D. Karande, and G. A. Patil, (2009), **Natural Language Database Interface for Selection of Data Using Grammar and Parsing**, World Academy of Science, Engineering and Technology.