A Review Paper on Hindi Language Graphical User Interface to Relational Database using NLP

Abhijeet R. Sontakke[1], Prof. Amit Pimpalkar[2]
[1,2] **Department of Computer Science and Technology**
[1,2] **G.H. Raisoni Academy of Engineering and Technology, Nagpur, India.**

Abstract:
   Data has been stored in the database and the databases are the major sources of information. Information is playing an important role in day-to-day life. This database technology has the major impact on the growing use of computer and internet. Database management system has been used for accessing, storing and retrieving data. However, database system is not understandable to each and every user because they are hard to use and understand. People with no knowledge of database language may find it difficult to access database. Therefore, there is need to find out the new technique and methods to access the database with the use of Natural Language Processing. Therefore this idea of using natural language instead of SQL triggered the development of a different type of processing method called Natural Language Interface to Database (NLIDB). Where user do not have any need to learn the formal language, they can give query in their native language .for the people who are comfortable with the Hindi language need this application to accept Hindi sentence as a query, process it and after execution provide result to the user in the same language which is nothing but the Hindi Language Interface to Database Management System.
Keywords: **DBMS, HLIDBMS, NLP, NLIDB, SQL.**

I. INTRODUCTION

   As the requirement of information and data is very important part of life. There are various sources of information but the major one is the databases. Database helps us to store, access and retrieve information. Any organization or industry is possible without the use of database. Each n every computer applications are dependable on database to access the information. For that it is necessary to have knowledge of formal query language like SQL but it is very difficult for everyone to learn and write SQL queries. To overcome this problem many researcher have brought out to use Natural Language (NL) i.e. English, Hindi, Marathi, Bengali, Arabic etc. in place of formal query language which can be perfect interface between an application of computer and non technical user.

This idea of using Natural Language has introduced the development of new type of processing method in database system. This new system can be named as (NLIDBs) i.e. Natural Language Interface to Database System. It is nothing but the communication channel between the user and the computer. Without any knowledge of any programming language, a user can act as a programmer. It will be very easy for a person to access data from database who has no knowledge of formal query language. NLIDB is basically the fields of NLP. A person who is not having any knowledge of database language may find it difficult to access database easily. Therefore SQL tutor will analyze the abilities of NLP to develop the product for people to interact with the database in simple English .because of these products there is revolution in extracting information from databases. Some of advantages and disadvantages of the system is given below.

We are going to develop a system for people who know Hindi language. User can access database using Hindi language and get the result in the same language. For example

सभी विधयर्थियो के नाम, अंक बताओ जिनका शहर नागपुर है. For our system we are going to use relational database as a test case.

II. LITERATURE SERVEY

   There is huge development in the area of NLIDB. Researchers are working on it from many years. Researchers like Androutsopoulos, G.D. Ritchie and P. Tanisch gives various architectures for NLIDB which is given below.

2.1 Architectures used by researchers for NLIDB
Architectures in the form of pattern matching systems, syntax based system, semantic grammar systems are developed by the researchers n that are explain below.

2.1.1 Pattern matching system

In pattern matching system patters n rules are given and that patterns and rules are fixed. The rules are, if input sentence or word is match with given pattern, the action has been taken and that actions are also mention in the database. But it is for some limited database and to the number of complexities of its pattern [1]. The advantage of this system is no parsing and module needed and system can be easily implemented. Some systems are working effectively but some would lead to be failed. SANVY is the best example of the pattern matching system [2][3].

2.1.2 Syntax based systems

In syntax based system user questions are analyzed syntactically i.e. it is parsed and the resulting syntactic tree is mapped to an expression in some database query language. One of the examples of syntax based system is LUNAR [4]. In this system grammar is nothing but the possible syntactic structure of the user's question. The advantage of this system is that it will give the detail information about the structure of the sentences.

2.1.3 Semantic grammar system

It is similar to the syntax based system. The query result is obtained by mapping the parse tree of sentences to a database query. The basic logic behind the semantic grammar system is parse the tree by removing unnecessary or combining the nodes together. Semantic grammar is used in PLANES [5] and LADDER [6].

2.2 Existing system which uses NLIDB

The best example of NLIDB system is LUNAR. That system comes in early 1973. LUNAR system is a system which answers the questions about rock samples brought back from the moon. LUNAR system is very effective which will handle 78% of request without any errors and that ratio riser up to 90% when the errors were corrected [5][7][8]. LADDER system was invented in 1978. It is first NLIDB for US Navy Ships. It uses the semantic grammar architecture concepts that interleave syntactic and semantic processing. LADDER was implemented in LISP. CHAT-80 was developing in 1980. It transforms English questions into prolog expressions which will be evaluated into prolog databases. In the same year the INTELLECT system was develop for ROBOTS [8]. The ASK system was develop in 1983. It was working for the end user to teach the system new words and concepts at any point during the interaction. It is the information management system with its own database. It will interact with multiple external databases, electronic mail programs and other computer applications.

Another system called GINLIDB (General Interactive Natural Language Interface to Database). UML is used to design it and it become developed with the visual basic .net 2005 [9]. START (Syntactic Analysis using Reversible Transformation) natural language system was developed in 1993. It was world's first question answering communication system [10]. It has language dependant function like parsing, natural language annotation to represent the appropriate information to the user [11]. JUPITER was developed in 1997. It was for weather forecasting information for the people worldwide [12]. ITS i.e. intelligent tutor system was developed in 1998 named as SQL-Tutor, which helps the student for getting answers of their question from the databases and the information stored in it [13]. The TREC [14] i.e. text retrieval conference was developed by soo-min kim and his colleagues. It again the question answering system for students which uses the various NLP techniques as well as word net for finding candidate answer.

B. Sujata etal [15] introduced a concept that, the Structured Query Language (SQL) norms are been pursued in almost all languages for relational database management systems. However, not everybody is able to write SQL queries as they may not be aware of the structure of the database. So this has led to the development of Natural Language interface for databases. There is an overwhelming need for non-sophisticated users to query relational databases in their natural language instead of working with the syntax of SQL. As a result many natural language interfaces to databases have been developed, which provides different options for manipulating queries. The idea of using Natural Language instead of SQL has prompted the development of new type of processing called NLIDB. Mohil Dua etal [16] had implemented the system based on NLP which gives output on the basis of NLIDB and Hindi language interface to database management system that give the proper result for only select, update and delete queries.

III. PROPOSED SYSTEM

3.1 Problem statement

We are going to use the STUDENT database as a case study for developing Hindi Language Interface to Relational Database using NLP. To retrieve information from database the query will be asked in Hindi language. We are going to provide the facility of the entire basic query like select, create, update and delete. We are also working to provide the advance query operation such as order by queries i.e. ascending order and descending order and most

3394

important we are going to provide the functionality of aggregate functions such as MIN(), MAX(), SUM() and AVG(). User will give the input in Hindi language and able to see the result in same language. SQL query is also displayed on graphical user interface.

3.2 Objectives

To develop a system for people who know Hindi language. User can access database using Hindi language and to get the result in the same language. All other objectives are given bellow-

i. To design Hindi language interface to relational database.
ii. To design GUI in which the end user can input the query in Hindi Language.
iii. To develop a system that can handle Hindi query for the extraction of single or multiple columns from tables stored in databases.
iv. To develop a system that can perform operations like select, update, delete on relational databases.
v. To extent the system to execute queries that involve joining of tables.
vi. To execute queries that include aggregate functions like sum (), max (), min (), avg ().
vii. To execute queries like arranging the data from the database in ascending order and descending order.
viii. To work on semantic behavior of the independent query.
ix. Analysis on the basis of time complexity.

3.3 Methodology

To achieve the above objectives some methodology has been followed and it is given below

i. Study of some IEEE papers has been done which is dependable on NLIDBs developed by Language Technology Research centre (LTRC) at IIT Hyderabad.
ii. Create student database which will store information about student.
iii. Identify the nature of queries i.e. select, update, delete, create, order by queries and aggregation functions.
iv. Appropriate mapping of tokens with database values should be done by extracting table, columns information from input Hindi sentences.
v. With the help of stored values of databases generate SQL query by mapping input query.
vi. Execute the query n give output in Hindi Language.

3.4 Architecture of the system

Architecture of Hindi language interface to relational database using NLP is given and explained and given below.

This architecture is known as HLIDBMS i.e. Hindi Language Interface to Database management System. There are five important phases i.e. Tokenizer, mapper, Query generator, DBMS and Linguistic Component.
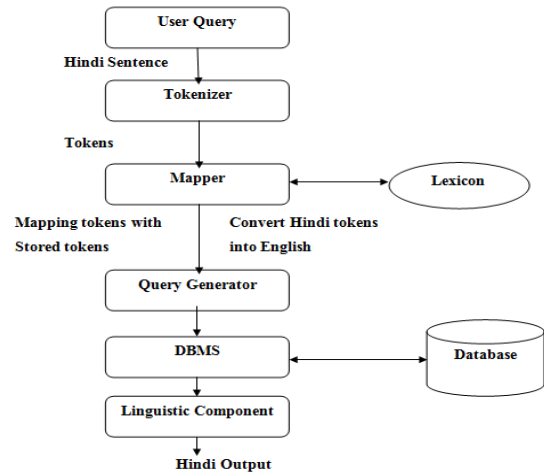

Figure: Architecture of HLIDBMS

In tokenize phase Hindi sentence is split into tokens. This is done with fact that all the tokens are separated by a space gap from each other. All the tokens which we get in this phase are stored in an array. Tokens are words of Hindi language. Token may be a table name, column name, condition, any value, command name, operation name or any non-useful word. To understand this; let the user query is as:

उन सभी विधार्थीयो का नाम, अंक बताओ जिनका शहर 'नागपुर' है This Hindi sentence has 11 tokens. First token is उन which is the starting of sentence. This is useless token which has no need in query formulation. Even in the absence of this type of tokens the query result will be unaffected. Likewise सभी is also non-useful token. Some tokens may be fields name as in the above query नाम and अंक are the field names. Input sentence also have token with table name विधार्थीयो. Therefore after this step we will have with all the tokens which the sentence is composed of.

Lexicon store all the Hindi tokens, their corresponding English word and type of token whether it is table name, column name, any value, operation, command or something else. Tokens which we extracted in above step are matched with the tokens stored in lexicon one by one. If it matches

then its corresponding English word is saved along with its type. This is the most important phase. All the useless tokens discarded in this phase only useful tokens are stored. After this step we will have with table name, field name (columns name), conditions, function etc. which will be further used to make SQL query. To formulate SQL query the beginning of this phase we are with table name, column name, condition and command etc. SQL is generated in this phase according to Hindi sentence. Execute query and display result to user the above get SQL query is executed and result of which in Hindi language is displayed to user on the interface.

The output is in the form of Hindi language and we are giving query also in Hindi language and processing of all this has been done by inner module as explained above. Some of the examples of expected input and expected outputs have been given bellow.

Suppose if user is giving query in Hindi language his query should be the format like given below in the table which represents the appropriate SQL query.

Hindi Query	Type of SQL Query
सभी विधयर्थियो के नाम, अंक बताओ जिनका शहर नागपुर है.	Select
गणेश की जन्मतिथि बताओ.	Select
उस विधयर्थी का नाम बताओ जिसे 70 के ऊपर अंक प्रास हुए है.	Select
सभी विधयर्थियो मे सेराकेश नाम के विधयर्थी का पता बताओ.	Select
जिस विधयर्थी का नाम उमेश, रोल नंबर 25 है उसका शहर मुंबई करो.	Update
सभी विधयर्थियो मे से नाम हटा दो जिसका रोल नंबर 19 है.	Delete
सभी विधयर्थियो के रोल नंबर चढ़ता क्रम मे लिखो.	order by query
सभी विधयर्थियो के रोल नंबर उतरता क्रम मे लिखो.	Order by query
सभी विधयर्थियो मे से किसके अंक जाड़ा है बताओ.	Aggregate query
सभी विधयर्थियो मे से किसके अंक कम है बताओ.	Aggregate query

IV. APPLICATIONS OF HLIDB

Hindi is highly spoken language in north and central India, Pakistan, Fiji, Mauritius. Approximately six hundred million people speak Hindi as their first or second language. Hindi is also a national language of India. Large number of e-governance applications uses databases. So to easily access the databases query given by user should be in Hindi language. For this there should be a system that accept the Hindi language, process it and generate the SQL query and gives the desired results in Hindi language only. Following are the some of the areas where the Hindi language interface can be applied.

- **Railways : Hindi language interface to railway database system**

Hindi is most common language used in India. So there is need to develop Hindi language interface to railway database system. The quarries of passenger such as reservations timings, cancellation of journey etc., are able to put in their native language. Passenger will give query in Hindi and he is able to get result in Hindi language only.

- **Agriculture : Hindi language interface to agriculture database system**

In India the whole population is dependable on agriculture. Farmers are generally not literate. They face lots of problems regarding irrigation, use of pesticides, time to reap the crop etc. Indian government has developed many systems to help farmer solving their queries. Data related to their queries is stored in databases but farmers is not that much literate to learn or know the SQL languages Hence the facility should be given so that the farmers can access the data from database it their own language so that the should utilize all the techniques n facilities related to the agriculture.

- **Weather forecasting and its related databases**

Weather report is necessary to each and every person. So providing weather report in Hindi is very much important. Farmers may easily get the weather forecasting report by asking query in Hindi. This can be done because of Hindi language interface to weather forecasting database system.

- **Legal matters : Hindi language interfaces to legal databases**

Lawyers' uses Hindi languages in their daily activities n hence they need to maintain their databases for clients. They just want a system which can process their request easily for them the Hindi

language interface to legal database system is very useful.

- **Employees : Hindi language interfaces to employee databases system**

Employees who are poor in query languages and English language, this Hindi language interface to employee database system helps them a lot

V. CONCLUSION

This Paper is on Hindi Language Interface to Relational Database using NLP accepts the query in Hindi Language and gives output in the same language. With the help of tokenizer, mapper, query generator, DBMS and linguistic component it is able to perform the appropriate operation and gives output in Hindi language which is easy to read and understand for the people who do not have any knowledge of SQL language or other query language. It is very much useful for non-technical person to retrieve data from database and get knowledge from it. With the help of this Hindi Language Interface to Relational Database using NLP we are able to perform all the operation such as select, update, delete, create and also the order by queries such as ascending order and descending order and aggregate functions such as min(), max(), sum(), avg() etc.

VI. REFERENCES

[1] Mrs. Neelu Nihalani. Dr. Sanjau Silkari and Dr. Mahesh Motawani, "Natural Language interface for Database: A brief Review", IJCSI International Journal of Computer Science Issues, vol. 8,no. 2, pp. 600-608, March 2011.

[2] T. Johanson, "Natural Language Computing – The Commercial Applications", the Knowledge Engineering Review, vol. 1, no. 11-23, 1984.

[3] Androutsopoulos, G.D. Ritchie, and P. Thanisch, Natural Language Interfaces to Databases - AnIntroduction, Journal of Natural Language Engineering 1 Part 1 (1995), 29–81.

[4] W. Woods, R. Kaplan, and B. Nash-Webber, "The lunar sciences natural language information system," 1972.

[5] D.L. Waltz, "An English Language Question Answering System for a large Relational Database", Communications of the ACM, pp. 526-539, 1978.

[6] G. Hendrix, E. Sacerdoti, D. Sagalowicz, and J. Slocum, "Developing a Natural Language Interface to Complex Data", ACM Transaction on Database system", communications of ACM, pp. 526-539,1878.

[7] W.Woods, "An experimental parsing system for transition net grammar in NLP", Algorithamic press, New York, USA, 1973.

[8]   L.R. Harris, "Experience with Intellect: Artificial Intelligence Technology Transfers", the AI magazine, pp. 43-50, 1984.

[9] Faraj A. EI-Mouadib, Zakaria S. Zubi, Ahmed A. Almagrous and Irdess S. EI-feghi, "Generic interactive natural Language Interface to Databases(GINLIDB)", International general of Computers, vol. 3, no. 3, 2009.

[10] "START natural Language Query Answering System"[online].Available: https://start.csail.mit.edu/

[11] M. Joshi, R. A. Akerkar, "Algorithm to Improve Performance of Natural Language Interface", International Journal of Computer Science and Applications, vol. 5, no. 2, pp. 52-68, 2008.

[12] V. Zue et al, "JUPITER: a telephone based conversational information for weather information", IEEE Transaction on Speech and Audio processing, vol. 8, no. 1, pp. 52-68,2008.

[13] Seymour Knowles and Tanja Mitrovic, "A Natural Language interface for SQL-tutor", Nov 5, 1999.

[14] Rajendra Akerkar and Pritti Sajja, "Natural Language Interface: Question Answering System" in Knowledge based System, Johnes and Bartlett Learning, pp. 333-344, Aug. 2009.

[15] B. Sujata, S. Viswanadha Raju and Humera Shaziya, "A Servey of Natural Language Interface to Database Management System" International Journal of Science and Advance Technology", vol.2, no. 6,june 2012 .

[16] Mohit dua, Sandeep Kumar, Zorawar Singh Virak, "Hindi Language Graphical User interface to Database Management System" 12th International Conference of Machine Learning and Application 2013.

BIOGRAPHIES

Abhijeet Ramrao Sontakke, pursuing M-tech from G.H. Raisoni academy of engineering and technology Nagpur, Maharashtra, India. Field of interest is NLP (Natural Language Processing) and Object Oriented Systems.

Prof. Amit Pimpalkar, Assistant Professor at G.H. Raisoni Academy of Engineering & Technology, Nagpur

Area of intrest is Natural Language Processing, Data Mining, AI and Image Processing.