Punjabi Natural Language Interface to Database

Jigha Kapoor
Computer Science Engineering
Bahra Group of Institutes, Behdpura, Patiala, Punjab, India

Harpreet Kaur
Assistant Professor
Department of Computer Science Engineering
Bahra Group of Institutes, Behdpura, Patiala, Punjab, India

Abstract- Natural Language Interface to Database Systems is a type of communication channel between the user and the computer. The user no longer needs to learn any SQL queries. Without the knowledge of any programming language, a user can act as a programmer. It becomes very easy for a person to access data from database who has no knowledge of formal query language. This system is mainly designed for users those are comfortable with the Punjabi language, they use in their daily life. In this paper, we used the Punjabi language interface for patient query database to query the database in Punjabi language. The system is proposed in .NET.

Keywords –NLP, Punjabi, Natural language interface, SQL, Hospital.

I. INTRODUCTION

In the digital era amount of data is increasing day by day. As the requirement of information is essential part of our life [1]. There are numerous sources of information, but the major one is database. The data are easily accessible if it is stored in a structured way or stored in a database. Database helps us to store, access and retrieve information. Databases Management and technologies of database are having a major impact on the increasing use of computer [2]. No organization or industry is possible without the use of database. Each and every computer based application need to access information from database that requires knowledge of formal query language like SQL. But it is not possible for everyone to learn or write SQL queries. To overturn this problem many researchers have brought out to use Natural Language (NL) i.e. Punjabi, Arabic, English, Bengali etc. in place of formal query language which can be a perfect interface between an application of computer and nontechnical user. This idea of using NL has induced the development of new sort of processing method in database systems. This new system can be named as Natural Language Interface to Database Systems (NLIDBs). It is a type of communication channel between the user and the computer. The user no longer needs to learn any SQL queries.

Without the knowledge of any programming language, a user can act as a programmer. No such hectic queries are required from the user, by the system.

1.1 HISTORY OF PUNJABI LANGUAGE

Punjabi, often spelled Panjabi, belong to the Indic group of the Indo-European family of Languages [3]. It is spoken in Punjab, neighboring states of Haryana and Himachal Pradesh. In India it is the official Language of Punjab state. In addition about 25 percent of the public living in the New Delhi metropolitan area speaks Punjabi in daily life [4]. The Punjabi language is closely related with the Sikh religion. Its alphabet, recognized as Gurmukhi, was the vehicle for recording the teachings of the Sikh gurus. It was invented by the second of the gurus in the 16th century. The word "Gurmukhi" means "Guru's mouth."

1.2 ARCHITECTURE OF NLIDB

There are four phases in the Architecture of NLIDB (Natural Language Interface to Database Systems) system as shown in figure 1. The four phase are- Token filters, Parser, Query creator and DBMS [1]. Output generated by tipper phase is used by lower phase as it's Input. Descriptions of all phases are given below:

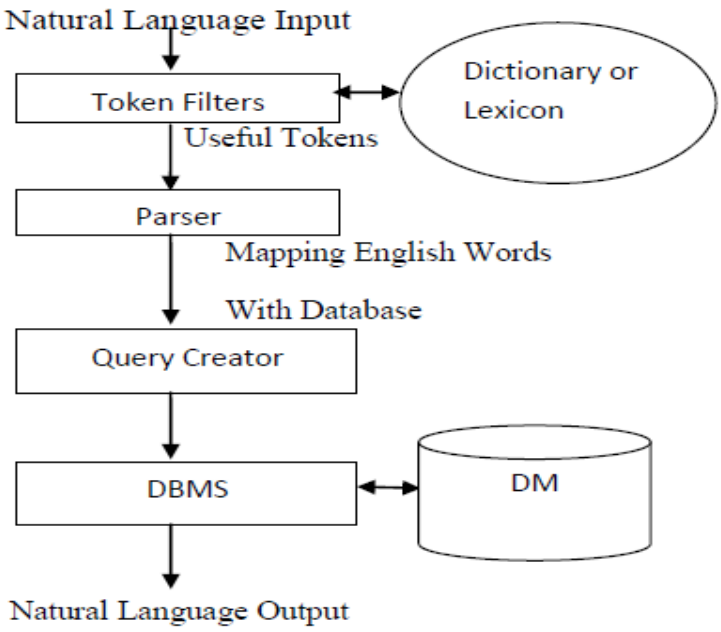

Figure 1. Architecture

1.2.1 Token Filter

This phase takes input from the user in Natural language. It then splits the sentence of Natural Language into tokens. All the tokens must be separated by comma, space etc. from each other. These tokens are then stored in an army. Tokens may represent name of a table, column, row, command, operator or it may be any value or any non-useful word. These all

the tokens are output of this phase. Some tokens may be conditions or values etc. This phase's output is given to the next phase.

1.2.2 Parser

This phase is most important and takes the input from the Token Filter phase. All the tokens, their corresponding English word and token type are stored lexicon or dictionary. These tokens may be name of a column, table or it may be any value, operation, command or something else. The tokens extracted by the upper phase are then matched one by one with the tokens stored in lexicon. If the match found then its corresponding English word with its type are saved. The tokens which are found useless are discarded in this phase. Only the useful tokens are kept. Now, we have with name of the table, attribute, conditions, and functions etc. that are used further to frame SQL query. The output of Token Filter phase is its input and it converts the tokens (that it takes from the above phase) into SQL quay. Only those tokens are stored in lexicon that is enough to create an SQL query.

For every token of natural language its corresponding English word and its type is also stored. The token stored may be a table name, column, function, condition etc.

1.2.3 Query Creator

This phase takes input from Parser phase. It takes name of a table, column, conditions, commend and functions. SQL query is created by this phase with the help of all these tokens.

1.2.4 Database Management System

In this phase SQL query is executed and the result is even to the linguistic component. This component then converts the result into user's own Natural Language.

1.3 PUNJABI LANGUAGE INTERFACE TO DATABASE

Large number of e-governance applications use database. So, to use such database applications with ease, people who are more comfortable with Punjabi language, require these applications to accept a simple sentence in Punjabi, and process it to generate a SQL query, which is further executed on the database to produce the results. Therefore, any interface in Punjabi language will be an asset to these people. With the help of Punjabi interface, they will be able to make use of database applications very well.

II.RELATED WORK

Hanane Bais et al. [5] presented the Architecture and the implementation of a generic natural language interface based on machine learning approach for a relational database. The advantage of this interface is that it functions independently of the database domain and automatically improves through experience its knowledge base.

Saravjeet Kaur et al. [6] proposed 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 user’s 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 proposed the architecture for translating English Query into SQL.

Wiqas Ghai et al. [7] discussed that Punjabi, Hindi, Marathi, Gujarati, Sindhi, Bengali, Nepali, Sinhala, Oriya, Assamese, Urdu are prominent members of the family of Indo-Aryan languages. These languages are mainly spoken in India, Pakistan, Bangladesh, Nepal, Sri Lanka and Maldive Islands. All these languages contain huge diversity of phonetic content. In the last two decades, few researchers have worked for the development of Automatic Speech Recognition Systems for most of these languages in such a way that development of this technology can reach at par with the research work which has been done and is being done for the different languages in the rest of the world. Punjabi is the 10th most widely spoken language in the world for which no considerable work has been done in this area of automatic speech recognition. Being a member of Indo-Aryan languages family and a language rich in literature, Punjabi language deserves attention in this highly growing field of Automatic speech recognition. In this paper, the efforts made by various researchers to develop automatic speech recognition systems for most of the Indo-Aryan languages, have been analysed and then their applicability to Punjabi language has been discussed so that a concrete work can be initiated for Punjabi language.

Manu Bansal et al. [8] discussed that the term data mining has been the oldest yet one of the interesting buzzwords. Many organizations often underutilize their already existing databases. There is a need to mine information and interesting patterns from these databases. The focus of the current research is to apply data mining on a library management system. Data mining is usually done on a data warehouse or a data mart. It incurs various cost factors like software, hardware, maintenance and experts. The objective here is to study how the real-time data stored in database can be turned informative without setting up a separate data warehouse. The main emphasize is on understanding the problem perspective, competing objectives and constraints and generating a model for information extraction from the real-time library database using ARM (Association Rule Mining) mining technique. As SQL (Structured Query Language) can also be used for mining data instead of using specialized data mining algorithm, the study also compares SQL based mining with ARM. The results shows that association rule mining performs better than SQL based mining as type of pattern to be extracted can be controlled much effectively in ARM as compared to SQL because of the parameters (support and count) used in the data mining algorithm. Algorithms are implemented using SQL and MATLAB (Matrix Laboratory) Tool - ARMADA.

Ravinder Kumar et al. [9] discussed that we require information in our daily life. One of the major sources of information is database. Almost all applications need to retrieve information from database that requires knowledge of database languages like SQL. To write SQL query one need to have knowledge of formal query language. Therefore everybody is not able to write SQL queries. To override the complexity many research have

turned out to use Natural Language (NL) i.e. English, French, Tamil, Arabic, Hindi, Punjabi etc. instead of SQL. The idea of using NL has prompted the development of new type of processing method called Natural Language Interface to Database systems (NLIDB). Lots of works have been already done in natural language processing but in Hindi and other Indian languages have scope of research and improvement. So for this intention we are trying to design a tool and database that helps the farmers to solve out the different queries requested in their native language (Hindi) related to agriculture. To design the tool that fulfills the objectives of our work we will use Java swings as front end and for storing the data we will use MySQL 5.0 as backend.

Priyanka Arora et al. [1] discussed that database Management Systems have been used extensively for accessing, storing and retrieving data. However, database systems are not understandable to every user because they are hard to use. A plethora of e-governance applications railways, billings, agriculture, banks etc. use databases. Some users face difficulty in using these database systems because they do not have knowledge of the languages used in this system. So, they want a system that accepts a Hindi sentence as s query and after processing it, execute it and provides the output in the same language. Then the users have no need to learn any low-level languages those are hard to learn, use in databases such as SQL.

Rohini B. Kokare et al. [10] discussed that natural language query builder interface retrieves the required data from database when query is given in natural language. To retrieve the correct data from database, the user should have sufficient technical knowledge of Structured Query Language (SQL) statements. Natural Language Query Builder Interface (NLQBI) will solve this problem. In natural language parsing, getting highly accurate syntactic analysis is a crucial step. Parsing of natural languages can be seen as the process of mapping an input string or a sentence to its syntactic representation. One of the parsing techniques is dependency parsing. Dependency parsing focuses on relations between words which resolve ambiguity. Most of the recent efficient algorithms for dependency parsing work by factoring the dependency trees. Graph based dependency parsing models are prevalent in dependency parsing because of their state-of-art accuracy and efficiency. This paper covers some recent developments in NLQBI systems and survey on dependency parsing techniques.

III. RESULTS

Punjabi Language Interface to database system provides an interface to the user which helps him/her to query the database in his/her natural language. In this work only Punjabi language is used as a mean for providing inputs. In this system we consider a database SQL Server and user created tables are used. A system is developed that eliminates the problem of normal user to interact with database with rigid language SQL. The users are able to access information's by issuing query in simple Punjabi language.

The main problem is to design a Natural Language Interface of Database for Punjabi language using patient query database. The query is asked in the Punjabi language for

retrieving the relevant information from the database. The format of the queries asked by the user must be simple, not complex. Hence, there is need of developing a Punjabi language interface for patient query database to query the database in Punjabi language. For the implementation of algorithm we have used C#.net as Frontend and SQL server as backend.

3.1 PATIENT ENQUIRY FORM

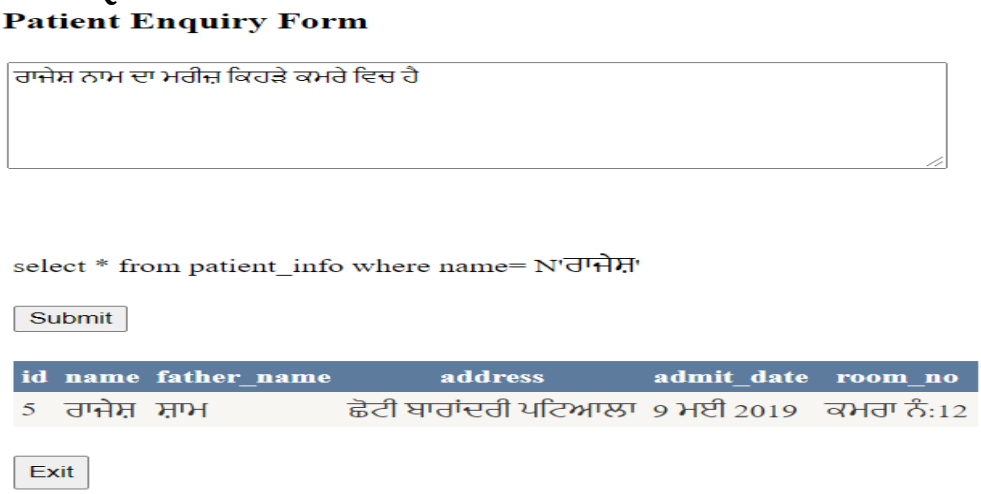

Figure 2. Patient Enquiry form

User can search the admit date, room no etc. of the patient by entering in Punjabi language in this form as shown in figure 2. Punjabi text is converted into SQL query and fetches the results from the database. Similarly, user can also search the entire information of the doctor.

In this form, user can also fix the appointment with the doctor as shown in figure 3.

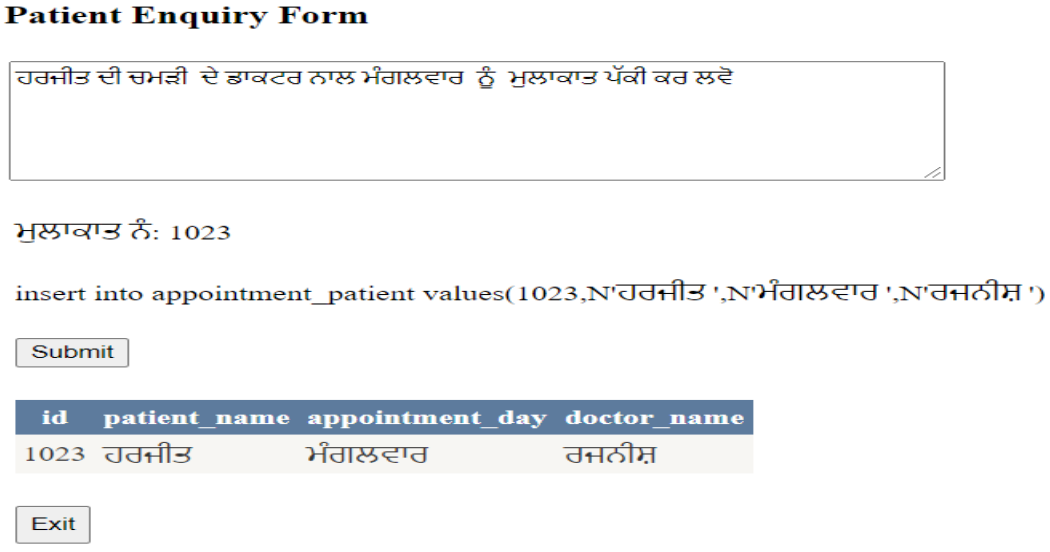

Figure 3. Patient enquiry form for fixing the appointment

IV.CONCLUSION

Punjabi Language Interface to Database is a form of computer-human interface. In the proposed system, user entered the input query related to patient or doctor in the Punjabi language and got the output results in the same language. The results showed that our software is correct and handles the SQL Queries without any problem.

V. REFERENCE

[1] Priyanka Arora ,Punet Goswami, "An Efficient Hindi Language Interface using Relational Databases", International Journal of Research in Computer and Communication Technology, Vol. 3, Issue 6, June – 2014.

[2] Garg, P., Singh, R. & Goel, M. B. (2020). A Novel Approach to Implement Natural Language Interface for Database in Punjabi Language. Journal of Xidian University, Vol. 14, No. 4, pp. 1230-1238.

[3] Deepjot Kaur, Navjot Kaur, "Implementation of DJ Rule Based Algorithm for DhuniVishleshan of Compound Punjabi Words", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 7, July 2013.

[4] Garg, P., Marwaha, A., Goel, M. B.(2020). Identification and Classification of Reduplication Words in Punjabi Language. International Journal of Scientific & Technology Research Vol. 9, No. 6, pp. 532-537.

[5] Hanane Bais, Mustapha Machkour, Lahcen Koutti, "Querying Database using a universal Natural Language Interface Based on Machine Learning", IEEE, 2016.

[6] Saravjeet Kaur, Rashmeet Singh Bali, "SQL Generation and Execution from Natural Language Processing", International Journal of Computing & Business Research, Proceedings of 'I-Society 2012' at GKU, Talwandi Sabo Bathinda (Punjab).

[7] Wiqas Ghai, Navdeep Singh, "Analysis of Automatic Speech Recognition Systems for Indo-Aryan Languages: Punjabi A Case Study", International Journal of Soft Computing and Engineering (IJSCE), Vol. 2, Issue-1, March 2012.

[8] Manu Bansal and Mandeep Kaur, "Analysis and Comparison of Data Mining Tools Using Case Study of Library Management System", International Journal of Information and Electronics Engineering, Vol. 3, No. 5, September 2013.

[9] Ravinder Kumar, "A Convenient Hindi Language Interface to Database for Agriculture Based Queries", International Indexed & Referred Research Journal, Nov-2014.

[10] Rohini B. Kokare & Kirti H. Wanjale, "A Survey of Natural Language Query Builder Interface for Structured Databases using Dependency Parsing", International Journal of Computer Applications,Vol. 107 – No 5, December 2014.