Natural Language Interface to Database

M. Humera Khanam* and V. Bhargavi†

†CSE Department, S.V.University, Tirupati, Andhra Pradesh, India

Accepted 10 May 2015, Available online 15 May 2015, Vol.5, No.3 (June 2015)

Abstract

A Natural Language Interface to Database (NLIDB) is a system that allows the user to access information stored in a database by typing request in some natural language (Telugu) can call it as Telugu Language Interface to Database (TLIDB). The main objective is to develop such an automated user interface so that system capable to give political information in Telugu Language. User can type questions in Telugu system will generate exact answers in Telugu so, this type of systems is very useful to native language speakers. Same question can be questioned in different ways but the meaning and answer is same in Telugu Language. This automatic system will give the best answer even though user pose questions different manner. TLIDB accepts Telugu query and generate appropriate Structured Query which helps to get the data from database then the resultant data is arranged in a Natural Language Answer and displayed to the user. Here, system generates exact and accurate answers. The accurate answers will be very useful and time saving. It is really helpful for the users who are using small screen devices, since in those devices it is very hard to find answers in a web page which contains lots of irrelevant content. Election information is useful to the users who are preparing for civil exams and school children’s.

Keywords: Question answering, natural language interface, database, NLIDB, SQL Statements.

1. Introduction

The process of Question Answering system is a technique of Information Extraction and Information Retrieval. Information Retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Information Extraction (IE) is the task of automatically extracting structured information from unstructured documents. IE is more about extracting from set of documents or information.

Most of the question answering systems are the process of retrieving appropriate answers for the user queries typed in natural language. Language is a medium of communication used by humans to express their views, ideas and emotions. Humans can able to learn new concepts and express their views is so natural but it is difficult to find how to process this language. Natural Language processing is a data driven empirical science. Natural language processing systems are built by training language independent and generic machine learning algorithms on large scale language data. Natural language processing (NLP) is an interpretation of language between human and machine. Natural language processing is so difficult because human language is complex and ambiguous. Language understanding needs contextual and general language apart from linguistic knowledge. Representing all this knowledge computationally is the challenging topic in NLP.

NLIDB systems are divided in to two sub components: 1. Linguistic component. 2. Database component. Linguistic Component handles input as natural language, translate it into formal query and generate output in natural language from the result which comes after execution of query. Database Component performs traditional database management functions. A knowledge base composed of a number of tables that store natural language words and their corresponding mapping to formal objects that are used to identify query frame. These tables can have entries of relations name; attribute names etc. questions entered in natural language translated into a statement with the help of parser which tokenize the input. Then a formal query is formed by mapping these tokens into knowledge base, which is executed and the result in natural language is given to the user.

There are so many approaches to develop NLIDB systems: 1. Pattern Matching Technique. 2. Syntax based approach. 3. Semantic Grammar based approach. 4. Intermediate Representation Languages. The proposed system is a Template Based Pattern Matching technique. These systems answers the user Natural Language query based on some matched patterns from the database. These are often managed to come up with some reasonable answer, even if the input is out
of the range of sentences the patterns were designed to handle.

2. System Architecture

In this pattern matching technique, the user enters the input as a query in his/her natural Language i.e., Telugu. Now, query statement is broken in to tokens, collects those Tokens and then use knowledge base to identify Keywords. The structured query format is selected based on tokens and the keywords in the input query statement which is useful to identify the minimum requirements that it can match with our database in order to have an accurate results. Each structured Query format is combined with a SQL generation process by using tokens. SQL statement is executed and extracts the result from the database. The resultant natural language answer is submitted to the user interface.

The Election QA system Architecture is shown in figure 1. User interface is the GUI visible to the user and can type questions regarding election information in Telugu. The input query is broken down in to tokens and identifies keywords from those tokens. Using predefined pattern matching templates, the keywords are filled in the slots of query templates. The SQL statement is executed and the result from the database is arranged in a structured answer format and finally, Natural Language answer is submitted to user interface using Template based approach.

3. Related Work

NLP researches have been working on Question Answering system since 1970’s with the systems like, BASEBALL (Green et al., 1971), it provides answers to the questions about the American Baseball. Peoples have questions and they need answers not documents. Automatic Question answering will definitely be a significant advance in the state of art information retrieval technology. Early QA SANYV is best natural language processing system uses pattern matching technique. The main advantage of pattern matching approach is simplicity.

In Syntax based systems, the input query is processed syntactically and generates parse tree. The resultant parse tree is mapped to query language to retrieve information from the database. This type of approach is used to domain specific database systems. LUNARis best example for this technique. The Geologists use the LUNAR system to ask questions about moon rocks.

LADDER system is good example for semantic approach. SANYV, LUNAR, LADDER all these QA system uses English as source and destination Language.

In the year 2006, dialogue based question answering system for Telugu language has been developed. In this system user can ask questions through speech and get the relevant answer through sounds. Here, Speech is the source and destination language.

A Hindi Question Answering System “Prashnottar”. In this system, input given by the user is HINDI language then system translates Hindi to WX format and searches on database. Finally, the resultant answer is translated from WX notation to Hindi language.

4 Design of Election Information System

The key role of election system lies in designing election database and Knowledge base.

Database means storing information in such a way that the information can be retrieved. Consider Relational database, information is stored in tables in the form of rows and columns. In election information system, database contains fields like constituency name, elected person, votes, party, symbol, state, Parliament, assembly are the relations in the election database. The political information like elected persons of MLA’s, MP’s, ministers and other information is stored in election database in the form of rows and columns. Data independence is the advantage of Relational model.

Election question answering system maintains knowledge base it is quite obvious in restricted domain. It greatly improves disambiguation and parsing. The Election QA system provides a user interface which is visible to user who can type their input query in Telugu. Whenever pupil types their input query the program will generate the tokens from the given input query and collects the keywords by applying stemming. The program will have predefined templates generated by the programmer, if user input query matched with the templates then the corresponding SQL query will execute and returns the result to the user interface. For example: if the user input queries will be,
“Who is Tirupati MLA?” The corresponding SQL query is,
SELECT name FROM assembly WHERE
constituency = "Tirupati";

The above SQL statement will give answer to the user using two tables such as, assembly, parliament using condition specified in WHERE clause.
If the input query,

Who is Chandragiri MLA? The corresponding SQL statement is,
SELECT name FROM assembly WHERE constituency = "Chandragiri" AND party = "";

The Election QA system will returns the result,

Who is Chandragiri MLA? Suppose if user requires information which is stored in multiple columns. For ex:

Who is Tirupati MLA? In this input query, user needs information about elected person and party using SELECT clause it is so simple to get the result stored in multiple columns just by specifying their field names after SELECT clause. The SQL statement for the above input query is SELECT name, party FROM assembly WHERE constituency = "Tirupati";

The natural language answer generated by the system for the above query is

Who is Tirupati MLA? If user needs information about how many constituencies are there in Andhra Pradesh? Here, the special operator \* is used. For ex: The Input query

How many constituencies are there in Andhra Pradesh? The SQL statement for this query is SELECT count (*) FROM assembly WHERE constituency in Andhra Pradesh; Here, COUNT function is used to return the number. Finally, the natural language answer generated by the system is

How many constituencies are there in Andhra Pradesh?

5 Experimental Results

Experiments are conducted to the automatic election QA system by checking with different models of queries as shown in below table.
Table 1: Different models of queries

<table>
<thead>
<tr>
<th>Natural Language Question</th>
<th>SQL Statement</th>
<th>Natural Language Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>తిరుపతి అసంబ్లి అభ్ెర్థి ఎవరు?</td>
<td>SELECT name FROM assembly WHERE constituency='తిరుపతి';</td>
<td>తిరుపతి అసంబ్లి అభ్ెర్థి.</td>
</tr>
<tr>
<td>చితత ిర్జిల్ో ఎనిి అసంబ్లి స్గు నాల్ు ఉనియి?</td>
<td>SELECT count(*) FROM assembly WHERE district='చితత ిర్జిల్ో';</td>
<td>చితత ిర్జిల్ో 14</td>
</tr>
<tr>
<td>చంద్రగథర్థ ఎమ్యెల్యె ఎవరు, ఏపగర్టీ అభ్ెర్థి?</td>
<td>SELECT name, party FROM assembly WHERE constituency='చంద్రగథర్థ';</td>
<td>చంద్రగథర్థ ఎమ్యెల్యె చేవ్విర్ెడిి బాస్కర్రెడిి బాస్కర్రెడిి.</td>
</tr>
</tbody>
</table>

TLIDB for election information provides the user interface which is available to the users who can type their questions in Telugu and then click SEARCH button. System process and generates Natural language answer to user interface from the election database. The snapshot for user interface as shown in below fig 2. The snapshot for query processing as shown in below fig 3. The snapshot for answer generation as shown in below fig 4.

Fig 2: User Interface

Fig 3: Query processing

Fig 4: Answer Generation

After, answer is returned to the user window. If user has one more question to query the system CLEAR button is available. Whenever user click on it, then the fresh home page is available where user can type their question.

Conclusions

The Proposed Natural Language Interface to Database performs pattern matching technique, uses election database, accepts input in Telugu and generate SQL query using tokens and keywords from the input query and natural language answer is submitted to the user window. The system performs operation on single table, multiple tables using functions AND, Nested Queries.

The future extension of this system is to implement a Telugu Interface system using Syntax and semantic approaches. Telugu is highly inflected with morphology, develop such a stemming algorithms without inflecting to the answers. Election is a huge information system in future can design systems which can handle complex user queries as well as aggregation functions in database.

References

J.B.Humphreys,(1961), Effect of composition on the liquidus and eutectic temperature and on the eutectic point of cast irons, BCRAJ,19, 609-621.
A.Udroiu,(2002), The use of Thermal Analysis for Process Control of Ductile Iron, Seminarium Nova cast, Italy.