International Journal of Current Engineering and Technology
©2015 INPRESSCO®, All Rights Reserved

E-ISSN 2277 – 4106, P-ISSN 2347 – 5161
Available at http://inpressco.com/category/ijcet

Research Article

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

*Corresponding author: **M. Humera Khanam** 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.

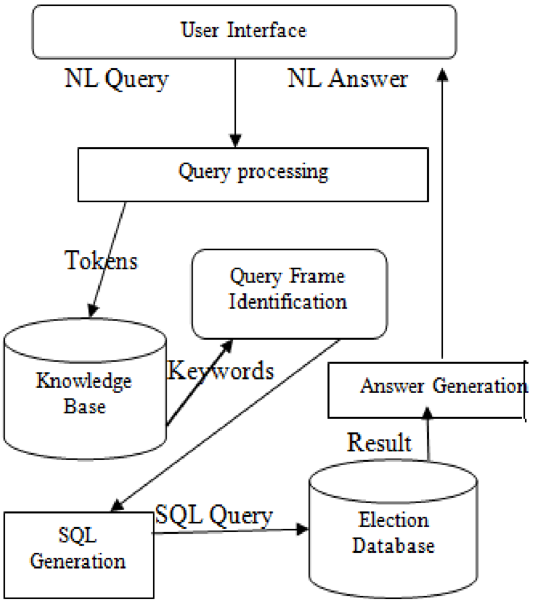

Fig 1: System Architecture

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 SANVY 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. SANVY, 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,

"తిరుపతి నేయోకవర్గానికి ఎమ్మెల్యే ఎవరు?"[tirupathi neyojakavarganiki mla evaru?]
"తిరుపతి ఎమ్మెల్యే ఎవరు?"
"తిరుపతి నియోజకవర్గానికి ఎమ్మెల్యే ఎన్నికయన అభ్యర్థి ఎవరు?"
"తిరుపతి నియోజకవర్గానికి టిడిపి అభ్యర్థిగా ఎన్నికయన వ్యక్తి ఎవరు?"

The above four questions have similar meaning i.e., "who is tirupati mla?" But the question framing is different. Here, తిరుపతి represents the constituency name, ఎవరు represents the Who, ఎన్నికయన, గెలిచినా represents the elected person. All these predefined knowledge is stored in the knowledge base in the form of look up table. The corresponding SQL statement for the above input queries will be

SELECT name FROM assembly WHERE constituency="తిరుపతి";

We store some of Telugu inflections like [ఏమ్యలేగా(mlaga),ఎన్నికయన(ennikayana),నియోజకవర్గానికి(neyojakavarganiki)] గా,యన,కి all these are morphological analysis of input query used to identify root words. After identification, system analyzes the whole input statement and gets the tokens such as, constituency, elected person, party, and some keywords. For ex: The input query,
చంద్రగిరి నుండి ఎన్నికయన వ్యక్తి ఎవరు? (chandragiri nundi ennikayana vykthi evaru?) Who is chandragiri mla? The input statement is parsed through spaces and applying stemming rules used to identify keywords which are helpful to frame correct query format. Here, in this example: చంద్రగిరి [constituency name], వ్యక్తి [person], ఎవరు [who]. The above input query represents the name of the person who elected as MLA form chandragiri constituency. Each structured query format is associated with SQL generation procedure. SQL is the language designed to use the relational databases. SQL statement for the above input query is
SELECT name FROM assembly WHERE constituency="చంద్రగిరి";

The above statement is executed and returns the natural language answer to the user.
[చంద్రగిరి నుండి ఎన్నికయన వ్యక్తి బాస్కరరెడ్డి]. Here, the result is collected from single table. The automatic election question answering system provides answers for factoid questions like (ఎవరు, ఎక్కడ,) for example:
వెంకటరమణ ఎక్కడ నుండి ఎమ్మెల్యేగా ఎన్నికయారు?
సతనపల్లి నేయోజకవరం ఎమ్మెల్యే ఎవరు?

In some cases, we need to use multiple tables to get an answer. Nested queries are used to retrieve answer from multiple tables. చిత్తూర్ జిల్లా ఎంపి మరియు ఎమ్మెల్యే ఎవరు?
The corresponding SQL query is,
SELECT name FROM assembly a, parliament WHERE a.constituency = p.constituency and constitutency= "చిత్తూర్";

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,
తిరుపతి నుండి ఎన్నికయన టిడిపి అభ్యర్థి ఎవరు? The corresponding SQL statement is,
SELECT name FROM assembly WHERE constituency = "తిరుపతి" AND party = "టిడిపి";

The Election QA system will returns the result,
తిరుపతి నుండి ఎన్నికయన టిడిపి అభ్యర్థి ఎం.సుగుణ.

Suppose if user requires information which is stored in multiple columns.For ex:
తిరుపతి నుండి ఎన్నికయన ఎమ్మెల్యే ఎవరు అతను ఎ పార్టీ అభ్యర్థి? 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="తిరుపతి";

The natural language answer generated by the system for the above query is
తిరుపతి ఎమ్మెల్యే వెంకటరమణ అతను టిడిపి అభ్యర్థి.

If user needs information about how many constituencies are there in Andhra Pradesh? Here, the special operator * is used. For ex: The Input query
ఆంధ్రప్రదేశ్ లో ఎన్ని నియోజకవర్గములు ఉనాయి? The SQL statement for this query is SELECT count (*) FROM assembly WHERE state="ఆంధ్రప్రదేశ్"; Here, COUNT function is used to return the number. Finally, the natural language answer generated by the system is
ఆంధ్రప్రదేశ్ లో 175 నియోజకవర్గములు ఉనాయి.

Once the SQL statement for an input query statement is generated, it is triggered on the database and the retrieved information is used to represent the answer. Each query frame has its corresponding Answer generator. We use template based answer generation method. Each template consists of several slots. Those slots are filled by the retrieved answer and the tokens generated from the query.

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 1Different models of queries

Natural Language Question	SQL Statement	Natural Language Answer
తిరుపతి అసెంబ్లీ అభ్యర్థి ఎవరు?	SELECT name FROM assembly WHERE constituency=' తిరుపతి' ;	తిరుపతి అసెంబ్లీ అభ్యర్థి సుగుణ.
చిత్తూర్ జిల్లలో ఎన్ని అసెంబ్లీ స్థానాలు ఉన్నాయి?	SELECT count(*) FROM assembly WHERE district= 'చిత్తూర్';	చిత్తూర్ జిల్లలో14 అసెంబ్లీ స్థానాలు ఉన్నాయి.
చంద్రగిరి ఎమ్మెల్యే ఎవరు, ఏ పార్టీ అభ్యర్థి?	SELECT name, party FROM assembly WHERE constituency= 'చంద్రగిరి';	చంద్రగిరి ఎమ్మెల్యే చేవ్విరెడ్డి బాస్కర్ రెడ్డి , వై.సి.పి పార్టీ అభ్యర్థి.

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.

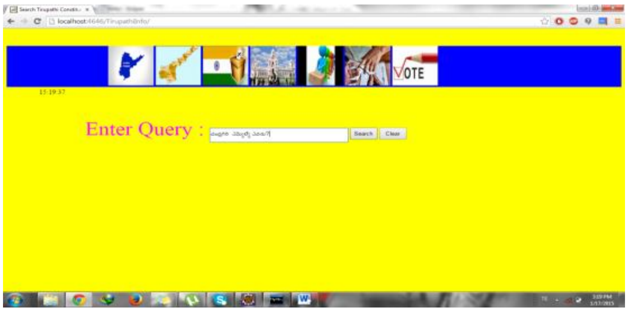

Fig 2: User Interface

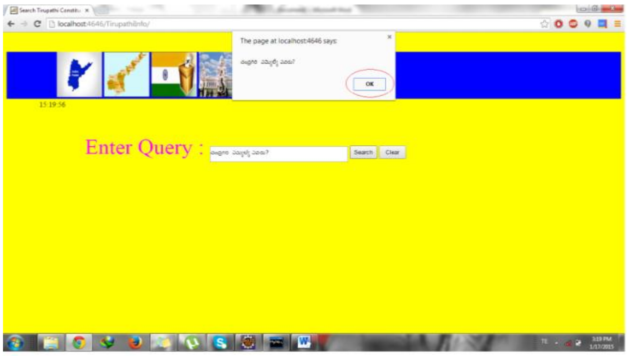
localhost:4646/TirupathiInfo/
The page at localhost:4646 says:
OK
15:19:56
Enter Query :
Search
Clear

Fig 3: Query processing

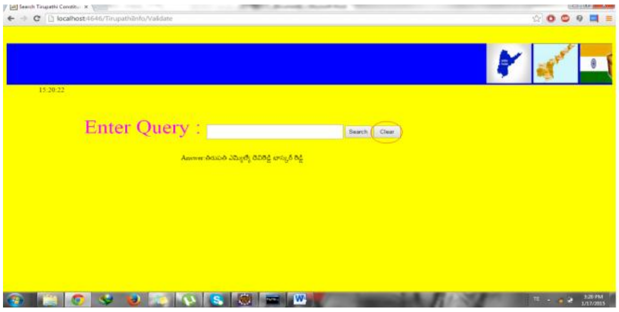

Fig 4: Answer Generation

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, *BCRIAJ*,19,609-621.

R.G.Warsinsk, (1975) Ford develops CE cooling curve computer, *Foundry M&T*,3,104-107

L.Backerud, K.Nilsson, N.Steen,(1975) The metallurgy of cast iron, *St.Saphorin, Swiitzerland Genrgi publishing company*,pp.625-637.

P.Zhu, R.W.Smith, (1995) Thermal analysis of nodular graphite cast iron, *AFS Transaction*, 103,601-609

C.Labrecque, M.Gagne,(1998), Interpretation of cooling curves of cast iron: A literature review, *AFS Transaction*,106,pp.83-90 M. Chisamera, I.Riposan, S. Stan, D. White, (2009), Influence of Residual Aluminum on Solidification Pattern of Ductile iron, *International Journal of Cast metals research,* vol.22,no.6, pp. 401-410.

I.Riposan, M.Chisamera, S. Stan, C. Gadarautanu, T. Skaland, (2003), Analysis of Cooling and Contraction Curves to Identify the Influence of Inoculants on Shrinkage behavior of Ductile Irons, *Keith Millis Symposium on Ductile Cast Iron*, pp.125-135.

A.Udroiu,(2002), The use of Thermal Analysis for Process Control of Ductile Iron, *Seminarium Nova cast, Italy*.

J. corneli, V.Ettinger, W. Baumgart, (2004), Thermal analysis ,an Unique Fingerprint of a melt ,*66th World Foundry Congress* 6-9 , pp. 743-756.

Seidu, S.O (2008). Influence of Inoculant's type on thermal analysis parameters of ductile irons, *4thinternaltion conference, Galati, Romania*, pp. 237-241.

M. Chisamera, S. Stan, I. Riposan, E. Stefan, G. Costache, (2007), Thermal analysis of Inoculated Grey Cast Irons, *UGALMAT,Galati, Technologiisi Materiale Avansate, University press*,Vol.1, pp.17-23.