Question Answering System for Election Database Using NLP

Dr.M.Humera Khanam[1], Mr.Md.A.Khudhus[2]
[1] Dept of Computer Science and Engineering, Sri Venkateswara University, Tirupati,
[2]BSNL, Tirupati,

ABSTRACT: In this paper, Question answering (QA) system for Election information using NLP has been described. The main goal is to extract election information in Telugu Language. User type input in Telugu, system will generate exact answers in Telugu so that this system is useful to technical as well as non-technical users. Same question can be framed in several ways but the meaning and answer is same in Telugu. This proposed system will give the answers smartly even though user frames queries in different manner. The election QA system is helpful to the pupil in different ways. This system needs an interface between user and database i.e., Telugu Language Interface to Database. It can be called as Natural Language interface to DB (NLIDB). By generating SQL Queries, answer can be collected from single Table or multiple tables. The accurate answers will be very useful and time saving. In this system we use pattern matching technique to extract the answer from database and produces answer in Telugu to the user.

KEY WORDS : Question answering, Natural Language interface to DB (NLIDB), SQL Statements, Election Data Base.Machine Learning.

1.INTRODUCTION

The process of question answering system is a technique in information extraction and information retrieval. Most of the question answering systems are the process of retrieving appropriate answers for the user queries typed in natural language. Question-Answering (QA) is defined as a task whereby an automate(such as computer)answers arbitrary questions formulated in natural Language. QA systemsare especially useful in situations in which user needs to know a specific piece of information and does not have the time or just does not want to read all the available documentation related to the search topic.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. Question answering (QA) system is an automated system capable to answer user questions in short and exactmanner. QA techniques can be classified into OPEN DOMAIN QA and Restricted-Domain QA, these two domains uses thesauriand lexicons in classifying documents and categorizing the questions. Open-domain question-answering deals with questions about nearly everything and can rely on general ontology, on the other hand, restricted-domain question-answering deals with questions under specific domain. Restricted Domain QA has a long history, beginning with systems working over databases (e.g., BASEBALL (Green et al., 1961), and LUNAR (woods et al., 1972)).The proposed question answering system for election information is restricted domain and it is monolingual language. i.e., both source and destination language is same (Telugu). Telugu is a source language for communication in AP and it belongs to Dravidian family.An automatic election QA system provides a user interface for users who can give their queries regarding election information. The system need to extract appropriate information from the database for this it needs an interface

between user and database. Lots of information is stored in database using tools like MS Access, Oracle and Others. Pupil who doesn't have an enough knowledge regarding Structured Query Language will face difficulty in handling and extracting useful information from the database. The proposed automated system is useful for non-technical and as well as technical users.

2.RELATED WORK

NLP researches have been working on Question Answering system since 1970's with the systems like, BASEBALL it provides answers to the questions about the American Baseball game. Automatic Question answering will definitely be a significant advance in the state of art information retrieval technology. Approaches that are used to develop Natural Language Interface to Database are

2.1. Pattern Matching Approach
2.2 Syntax based Approach
2.3 Semantic grammar based Approach
2.4 Intermediate Representation Languages

In Pattern matching Approach users input query directly matched to the database. Formal List Processor (FLP) is an early language for pattern matching based on LISP structure. SANVY is best natural language processing system uses pattern matching technique. The main advantage of pattern matching approach is it can be used with various data types (int ,float ,char),and its simplicity . In Syntax based systems, the users question is parsed (i.e. analysed syntactically) and the resulting parse tree directly mapped to expression in some database query language. And this approach uses a grammar that describes the possible syntactic structure of the user's questions. LUNAR is best example for this technique. The Geologists use the LUNAR system to ask questions about moon rocks. And this system uses an Augmented Transition Network parser. 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.

3. PROPOSED WORK

In the proposed Election QA system user types questions in Telugu and system understands and generate answers in Telugu. Answers are extracted from the database so that, the system needs an interface i.e., Telugu Language Interface to DatabaseUsing this system, user can query the database without by typing natural language questions and can access the requested information from the database. We call it as Natural Language Interface to Database (NLIDB). Pattern matching, syntactic, semantic these are the techniques in NLIDB. The current proposed Election QA system uses pattern matching technique and can access the data from single table or multiple tables using some operations like nested queries, conjunction and can also collect the data from related documents.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 the common keywords and then use knowledge base. The structured query format is selected based on tokens and the keywords in the input query statement 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 data. The fig 1 shows system architecture

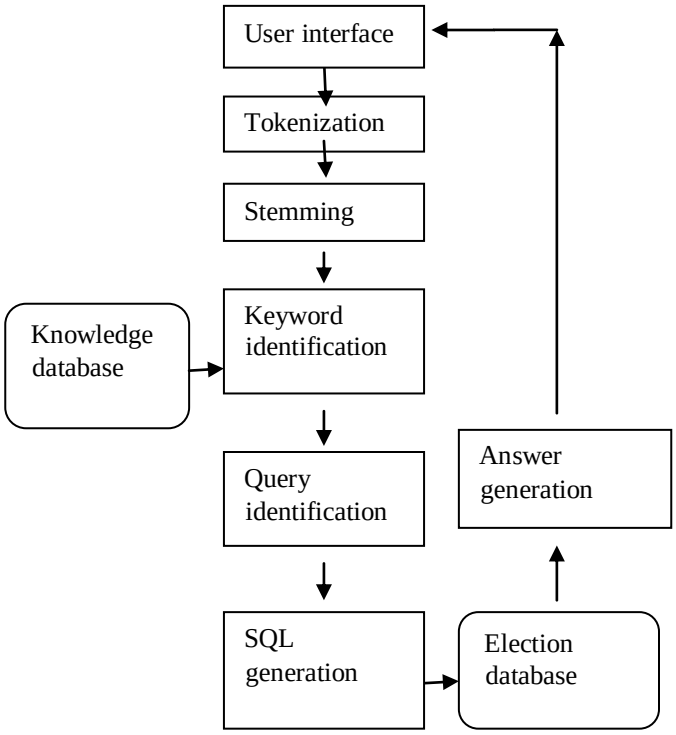

Fig.1 System architecture

4.DESIGN AND IMPLEMENTATION

Here we are discussing about analysis of source language which gives requirement to develop a question answering system for election database and presents system architecture that represents how user query the database through user interface, techniques to analyse source language and represents how to extract answers from the database.

Algorithm:

1. Input query in Telugu language
2.Tokenize the query
3.If inflected word do stem then GOTO step 4
4. If tokens==Look_up(Table_names)
   Else GOTO step 10
5.If token++Look_up(column_names)
   Else GOTO step 10
6. If token++LookUp(conditions)
   Else GOTO step 10
7.Generate SQL query
8.Execute SQL query
9.Present natural language answer to user
10. Invalid query-show error message.

4.1 User interface*:*

User interface is visible to the user who can request their queries and can see the results for their queries.

For Example: user types a query like

INPUT: నగిరినియోజకవర్గఎంఎల్ఏఎవరు?

OUTPUT:నగిరినియోజకవర్గఎంఎల్ఏఆర్.కేరోజా

4.2 Tokenization:

In this phase Telugu sentence is split into tokens. This is done with the fact that all tokens are separated by space from each other and these are stored in an array. Token maybe table name, column name, condition, or any value.

For Example: let the query be

Input:నగిరినియోజకవర్గఎం.ఎల్.ఏ?

The above query has 4 tokens

నగిరి,నియోజకవర్గ,ఎం.ఎల్.ఏ,?

4.3 Stemming

Stemming is the process of extracting the root word from the inflected word which is free from all inflectional and deviational endings. Generally it may following three changes to stem.

1.Removes a letter from stem and adds new suffix.

2.Adds the suffix without changing the stem.

3.Adds the suffix by changing the stem.

4.4 Keyword Identification

After parsing each token is searched in knowledge base until the word is found and these kept their types and semantic information are put in a list of tokens

For Example: పీలేరుఎంఎల్ఏఎవరు?

From the above tokens are identified areఎంఎల్ఏ[MLA] as table column,పీలేరు[pileru] as constituency name,ఎవరు[who] as keyword.

4.5 Query Frame Identification

During the analysis of the query the keywords in the input query are deleted in this step based on tokens and keywords we identify the appropriate query frame. Restricting the query domain and information resources the scope of the user request can be focused i.e, there is a finite number of expected question topics. Each expected question topic is defined under single query frame. Some examples are person name, party name, constituency name, votes.

For example: పీలేరుఎంఎల్ఏఎవరు?

The above query identified by the person name.

ఆంధ్రలోఎన్నిపార్లమెంట్స్థానాలున్నాయి?

The above question is identified by number of assemblies in Andhra.

4.6 SQL Generation

Structured query language is a widely used programming language for working with relational database. The abundant information available on the internet generates the need to store data in an organized

manner so that searching, retrieving, maintaining data will becomes easier.Database is a technology that stores the data in logical and organized manner.

1. Queries for selection of whole table
2. Queries for selection of certain columns
3. Queries for selection of certain rows from certain columns i.e., queries with"where" condition.

For Example: శ్రీకాఃలహస్తిఎంఎల్ఏఎవరు?

SQL Query:SELECT name FROM assembly WHERE constituency="శ్రీకాఃలహస్తి"

Example:కృష్ణజిల్లాలోఎన్ని అసెంబ్లినియోజకవర్గాలువున్నాయి?

SQL Query:SELECT count(*) FROMAssembly WHERE district="కృష్ణ"

4.7 Election Database Management System

Database means storing information in such a way that the information can be retrieved.It is structured and contains the information to provide the political/election results.It contains the data about the name of the elected persons,parties,constituencies, district, votes in the form of rows and columns.

4.8Answer Generation

SQL Query is executed and result of which in Telugu language is displayed to user on the interface. Using template based answer generation method. Each template consists of several slots and these filled by retrieved answer and tokens generated from the query. Table 1 shows the template.

అభ్యర్థిపేరు	నియోగకవర్గం	పార్టీ	వోట్లు	జిల్లా
తిప్పరెడ్డి	మదనపల్లి	వై.ఎస్.ఆర్.సి.పి	3625	చిత్తూర్
జి.శంకర్	తంబల్లేపల్లి	టి.డి.పి	4032	చిత్తూర్
మంజుల	పీలేరు	వై.సి.పి	7132	చిత్తూర్

For Example: కడపఎంపిఎవరు?

SELECT name FROM assembly WHERE constituency="కడప"

DBMS returns "అవినాష్"

కడపఎంపిఅవినాష్

5.EXPERIMENT AND RESULTS

In this election question answering system provides the user interface which is available to the users who gives input in the Telugu language and then click search button. System process and generates answer in the natural language to the user. The answer is returned to user if it is present in the database otherwise displays as error message and then clicks ok button.

Table:2

Natural language question	SQL statements	Natural language answer
చెవిరెడ్డిఏపార్టీఅభ్యర్థి?	SELECT party FROM assembly WHERE name ="చెవిరెడ్డి"	చెవిరెడ్డివై.సి.పి. అభ్యర్థి
నెల్లూరుఎంఎల్ఏఎవరుఏపార్టీకిచెందినఅభ్యర్థి?	SELECT name,party FROM assembly WHERE constituency ="నెల్లూరు	నెల్లూరుఎంఎల్ఏప్రభాకర్కాంగ్రెస్పార్టీకిచెందినఅభ్యర్థి
ఆంధ్రలోఎన్నిఅసెంబ్లిస్థానాలుఉన్నాయి?	SELECT count(*) FROM assembly WHERE state="ఆంధ్ర"	ఆంధ్రలో 175 అస్సెంబ్లీస్థానాలుఉన్నాయి

6.CONCLUSION

Here I concluded that the proposed automated Election Question Answering system for Telugu Language performs pattern matching technique uses election database, accepts input in Telugu and the input breaks into tokens based on these keyword identification will be done from the knowledge database and generate SQL query using tokens and keywords from the input query and answer in Telugu is submitted to the user window. The system performs operation on single table, multiple tables using functions AND, Nested Queries. The proposed system will retrieve answers to the factoid questions, description type questions regarding elections in Telugu Language. The future extension of this system is to implement a Telugu QA system using Syntax and semantic approaches.

Examples of Election QA System

Example 1:

User: కడపజిల్లాలోఎన్నిఅసెంబ్లినియోజకవర్గాలుఉన్నాయి? (Kadapa jillaloenni assembly neyojakavargamuluvunayi [How many constituencies are there in kadapa district])? The corresponding SQL statement generated by the system is, SELECT count (*) FROM assembly WHERE district=" కడప"కడపజిల్లాలో9 అసెంబ్లినియోజకవర్గాలుఉన్నాయి

Example 2:

User: రాజంపేటనియోజకవర్గంనుంచిఅసెంబ్లిసభ్యునిగాపోటీచేసినవై.ఎస్.ఆర్.సి.పిసభ్యుఎవరు? (rajampeta neyojakavargamnunchi assembly sabhunigapotichesina Y.S.R.CP abhyarthiperuenti)? The corresponding SQL query generated by the system is,

SELECT name FROM assembly WHERE constituency=" ర్గయచోట్"AND party=" వై.ఎస్.ఆర్.సిపి";

System:రాజంపేటనియోజకవర్గంనుంచిఅసెంబ్లిసభ్యునిగాపోటీచేసినవై.ఎస్.ఆర్.సి.పిసభ్యుధుఅవినాష్

Example 3:

User: కడపజిల్లాలోతొమ్మిదినియోజకవర్గాల్లోటి.డి,పిఎన్ని నియోజకవర్గాల్లుగెల్చుకుంది? (Kadapa jilalonithomidi neyojakavargalaloT.D.P ennineyojakavargamulugeluchukundi?) The corresponding SQL query generated by the system is, SELECT count (*) FROM assembly WHERE distruct=" కడప" AND party=" టి.డి.పి ";

System: కడపజిల్లాలోతొమ్మిదినియోజకవర్గాల్లోటి.డి,పి 6 నియోజకవర్గాల్లుగెల్చుకుంది

Example 4:

User: సర్వేపల్లినియోజకవర్గంఅసెంబ్లిస్థానానికిఏపార్టీగెలిచింది? (sarvepalli neyojakavargam assembly sthananiki e party gelichindi?) the corresponding SQL statement generated by the system is, SELECT party FROM assembly WHERE constituency="సర్వేపల్లి": System: సర్వేపల్లినియోజకవర్గంఅసెంబ్లిస్థానంకికాంగ్రెస్పార్టీగెలిచింది.

REFERENCES

1. W.A. woods R.M. Kaplan and B.N Webber " The LUNAR Sciences Natural Language Information Systems" Cambridge 1972.
2. Rami Reddy Nandi Reddy, SivajiBandyopadhyay "Dialogue based Question Answering System in Telugu" published at EACL 2006 Workshop on Multilingual Question Answering - MLQA06.
3. D. Ramesh, sureshkumar "Telugu Language Interface to Database" published at International Journal of Advanced research in computer and communication Engineering, july 2013.
4. Poonamguptha, vishalgupta "A survey of existing question answering techniques for Indian Languages" published at Journal of Emerging Technologies in Web Intelligence, may 2014.
5. S. QUARTERONI, S. MANANDHAR " Designing an Interactive Open domain question answering system, 2007 at Cambridge university.
6. Aksharbharthi, vineetchaithanya, Rajeev sangal "Natural Language Processing A paninian perspective", department of computer science and engineering at Indian Institute of Technology Kanpur.
7. Poonam Gupta, Vishal Gupta "A survey of Text Question Answering Techniques.
8. David L.waltz "An English Question Answering System for a Relational Database at University of Illinois at urban-champaign.
9. Hoojung Chung, Young-In Song, Kyoung-Soo Han, Do-Sang Yoon, Joo-Young Lee, Hae-Chang Rim and Soo-Hong Kim. 2004. A Practical QA System in Restricted Domains.