RESEARCH ARTICLE OPEN ACCESS

Natural Language Interface for Data Base: A Case of Hindi Language

Keshav Niranjan [1] Sonia Yadav [2]
[1] Department of Computer Science,Pusa Institute of Technology
[2] Department of Computer Science, Deshbandhu College, University of Delhi
New Delhi - India

ABSTRACT
The paper presents a paradigm for developing the Hindi Language Interface (HLI) for databases. The proposed HLI system will consist of several modules such as Input, Tokenizer, Query Mapper, Query Builder, Query executer and Output modules. Each module is connected to the next module and gives the intermediate output as input to the next module. The paradigm has been applied to Hindi language which is the most spoken language in India. The goal is to replace the Structured Query Language (SQL) with Hindi for database communication and analysis. The proposed system is expected to be useful to novice, nontechnical Hindi users who have difficulty in learning complicated SQL commands.
Keywords: SQL, HLI, SQLG, TDIL, NLI

I. INTRODUCTION

Database is necessary for every business in digital era and manipulation on database can be possible by the knowledge of restricted English(SQL) but country like India which is multilingual country where dialect is changing after 50 kilometers, there is no single accepted language in India. Generally Indian languages have classified in five languages families Indo-Aryan (76.87 % speakers), Dravidian (20.82% speakers), Austro-Asiatic (1.11 % speakers), Tibeto-Burman (1% speakers) and Andmanese (0 % speakers) [1,4] and approximately 50% Indian population speaks Hindi. 96% of Indian population does not speak/use English. The Database usage therefore is restricted to the 4% population in India. The proposed system therefore will enable a sizeable population of India to use Database and analyze data.

II. LITERATURE REVIEW

During 1960 various efforts have been made for developing the natural language interface. Initial effort for evolution of NLI can be seen as BASEBALL. BASEBALL(Green et al., 1961) was the first natural language interface for database. Woods built natural language interface LUNAR for the geology based database in 1971. This system answers all the questions about the Apollo moon rocks for the NASA Manned Spacecraft Centre. Transformational Question Answering System (TQA) (Petrick, 1981; Plath, 1976) is anatural language interface to databases. TQA permits to users to pose their queries in English rather than in some formal language such as SQL. RENDEZVOUS (Codd, 1974; 1978) is prototyping based natural language interface. RENDEZ-VOUS is based on phrasal approach. This 'phrasal lexicon' is used bi-directionally[1]. PLANE (Waltz, 1975; 1978; Waltz and Goodman, 1977) is natural language interface for a large relational database of aircraft flight and maintenance data. EUFID (Burger, 1977; Templeton and Burger, 1983) is semantic parsing based NLI. The LADDER (Sacerdoti,1977; Hendrix et al., 1978) system was designed as a natural language interface to a database of information about US Navy ships. ROBOT (Harris, 1977; 1978) is first commercially available natural language processor for hierarchical database management system[2]. PHLIQA1 was developed at Eindhoven Netherlands's Philips Research Laboratories by the team of W. J. Bronnenberg, H. C. Bunt, S. P. J. Landsbergen, P. Medema, R. J. H. Scha, W. J. Schoenmakers and E. P. C. van Utteren.[3]CHAT-80 is prototype Natural Language System for world of geography. CHAT-80 database is data collection of oceans, rivers, seas,cities. IRUS (Bates and Bobrow, 1983; Bates, Moser & Stallard, 1986) uses the RUS parser. Later this system was converted into a commercial interface PARLANCE.Transportable English database Access Medium (Grosz,1983; 1984; Grosz et al., 1987) is a natural language interface for the relational database. DATALOG (Hafner, 1984; Hafner and Godden, 1985) interface consists of feature of extendibility and portability. ASK (Thompson and Thompson, 1985) uses the simple dialect of English for query. In 1994 the final prototype of EMIR (European Multilingual Information Retrieval) was released. From 2000 to 2004, DARPA, the US Defense Advanced Research Projects Agency, supported the TIDES programme for Translingual

Information Detection, Extraction and Summarization with the goal of "enabling people to find and interpret needed information, quickly and effectively, regardless of language or medium"[4]. Government of India launched Technology Development for Indian Language program(TDIL). TDIL decides the major and minor goal for Indian language Technology and provides the standard for language technology[5]. Microsoft Corporation has launched Hindi language interface for the Windows XP operating system[6].

III. HINDI LANGUAGE INTERFACE (HLI) ARCHITECTURE

Processes of HLI will be as follows:

3.1 Input

User will pose the query in Hindi language such as

कर्मचारीकेनाम दिखाओ

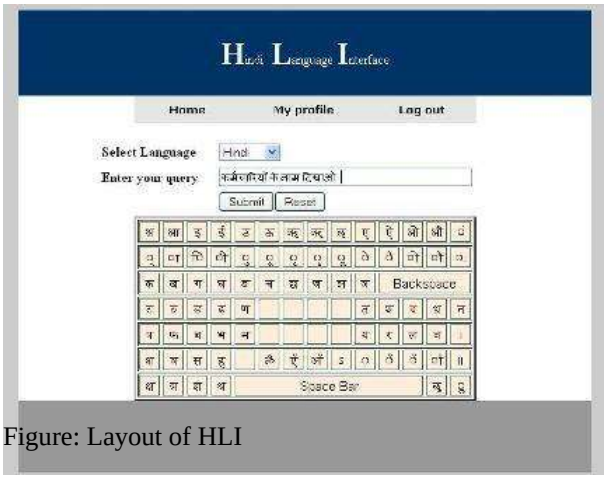


Figure: Layout of HLI

3.2 Tokenization

Tokenizer will tokenize the query on the basis of the white space parameter and query will produce the following tokens after tokenization

कर्मचारी | के | नाम | दिखाओ

t_1= कर्मचारी, t_2= के, t_3= नाम, t_4= दिखाओwhere t_1, t_2, t_3 and t_4 are tokens.

3.3 Lexicon:

Formulation of the query for the interface depend on the user's mental trait and commonly usable words in his culture and long time practice of user with the words. So the system will have the collection of possible words that user can use in query formulation and their corresponding SQL keywords. The query words lexicon will be as follows:

Table 1: Lexicons Table

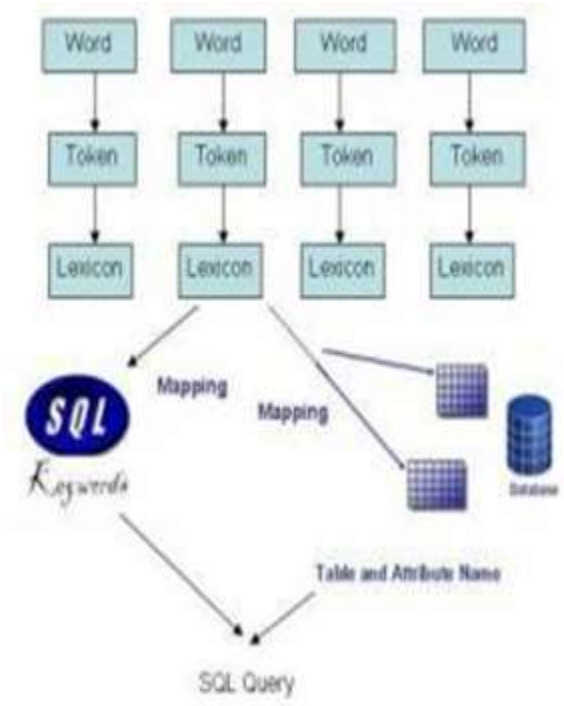
Select	दिखाओ,चुनना,चुनावकरना,चुनलेना,छाँटना,बीनना,कराना, प्रदर्शन,प्रदर्शनी,दिखावा,प्रकाशितकरना, छांटलेना, मार्गदर्शक, खोज,पतालगाना,प्राप्तकरना,ढूंढनिकालना,अन्वेषणकरना, अनुभवकरना,प्राप्तकरना, व्याख्याकरना, अधिग्रहण
From	से,द्वारा,कारणसे,आरंभकरके
Create	सर्जनकरना,बनावट,रचना,नमूना,स्वभाव,आकार,गठन,शकल,रूप,माडेल,निर्माणकरना,बनाना,तैयारकरना,रचना,बनना,बनादेना,बनालेना,कार्यकरना,रचनाकरना,उपजाना
Insert	प्रवेशकरना,दाखिलकरना,दर्जकरें,सम्मिलितकरें,अभिलेख, आलेख,लिपिबद्धकरना,रजिस्टरमेंलिखना, टांकना, अंकितकरना,पंजीकृतकरना
Update	आधुनिकतमबनाना,नवीनतमबनाना,रूपांतरितकरना, सुधारना,रूपबदलना,संशोधित,आधुनिकीकरण,नवीनीकरणकरना,नयाबनाना,आधुनिकसमयकेअनुसारबनाना,नवीनीकृत ,नवीकरणकरना,नयाकरना,नयाहोना
Delete	हटाना,मिटाना,काटना,क़लमखींचना,क़लमफेरना,काटदेना, बरबादकरना,तबाहकरना

3.4 SQL Query Generator

The Structured Query Language Generator(SQLG) consists of two modules - Mapper and Query builder. After the tokenization we match each token to the query lexicon. Since our objective is to identify the keywords in the input query which have SQL equivalents, the keyword identifier module will pick out SQL pertinent words to be matched in a lexicon of synonyms and equivalence. We can look into the lexicon table which has synonym of SQL pertinent word. A sample data is given below –

Mapper
Employees Discard Name Select
Query Builder
Select name from employee;

Layout for process of SQL query generator



3.5 Output

After the query generation, query will be passed to the SQL executer which will return the desired result in table form as given below in the figure.

Name
Ram
Shyam
Mohan

IV. ALGORITHM

System will follow the following algorithm for retrieving the data from the database.

1. Read Input statement S
2. for each word W_i from S do
   if(W_i=SelectKeywordSynonym)
   {
   FindTableName();
   FindAttributeName();
   }
   elseif(W_i=UpdateKeywordSynonym)
   {
   FindTableName();
   FindAttributeName();
   }
   elseif(W_i=DeleteKeywordSynonym)
   {
   FindTableName();
   }
6. end if
7. Display type of query Q
8. Formulate SQL query Q
9. Execute the Query
10. Display result

V. CONCLUSION

User has the following advantages from the NLI's:

1. Whatever native linguistic stuff is in user's mind, it is enough for manipulating the database because user can ask query simply in native language. Therefore knowledge of the SQL commands and structural knowledge of SQL query will not be necessary condition.
2. User doesn't require any technical training because knowledge of the Hindi is one acquires naturally from the culture.
3. It is easy to operate.
4. Recall and Precision will be higher than other methodologies, so its output will be more accurate. from all these advantage we can say that it is viable option for the novice and non technical users.

REFERENCES

[1] Niranjan,Keshav,2012,Language technology in India, Language in India,vol .12:4 p 179-187.
[2] Niranjan,Keshav,2011,Natural Language Interface for Database:Using Keyword Approach,Excel India Publisher,New Delhi,India.
[3] El-Mouadib ,Faraj A. ,and Zubi,Zakaria, Suliman and Almagrous ,Ahmed, A.,and El-Feghi, I.,2009," Interactive Natural Language Interface ",WSEAS Transactions On Computers, ISSN: 1109-2750, Issue 4, Volume 8.
[5] Jha, Girish Nath, 2003, Current *Trends in Indian languages Technology*, Language in India, Volume 3:12. Girish Nath,India's language diversity and resources of the future: challenges and opportunities, Special Center for Sanskrit Studies Jawaharlal Nehru University, New Delhi.
[6] Akerkar, Rajendra and Joshi, Manish,"Natural language interface using shallow parsing", International Journal of Computer Science and Applications,Vol. 5, No. 3, pp 70 - 90
[7] http://www.cdacindia.com/html/about/success/mantra.aspx
[8] http://www.cdacnoida.in
[9] http://www.indiareports.com/corporate/Guruji.aspx
[10] http://tdil.mit.gov.in/AboutUs.aspx
[11] http://tdil.mit.gov.in/Standards/ISCII.asp