Natural Language Database Interface For Select Sql query with Probabilistic Context Free Grammar

Mr.Sunil M.Jadhav [1] *Ms. Khushbu Doulani [2]* *Ms. Gitanjali B. Yadav[3]*

[1, 2, 3] Department of Computer Engineering,
Rajgad Dyanpeeth Technical Campus,Pune University,
Dhangwadi Pune 412205,Maharashtra, India.

[1]suniljadhav02@gmail.com, [2] khushidoulani@gmail.com , [3]gitanjali3014@gmail.com

Abstract : *A Natural Language Interface to a Database (NLIDB) is a system that allows the user to access information stored in a database by typing requests expressed in some natural language. (NLIDB) are systems that translate a natural language sentence into a database query NLDBI system including its probabilistic context free grammar, which can be used to construct the parse tree, an algorithm to calculate the probabilities. We specify the model for helping the user with queries depending up on probabilistic context free grammar (PCFG) to relational database.*

Keywords: ***NLDBI, Probabilistic Context Free Grammar, SQL Translator, Experimental Methodology***

1. NLDBI

NLDBI (Natural Language Database Interface) is a system that allows users to access a database in natural language and has been a popular field of study. NLDBI allows the users to access the database even though they doesn’t have the database dependent SQL Queries. User enters his query with the help of interface. As all the employees in an organization may not be aware of the SQL queries so the user cannot access the database content directly. The user who has the knowledge of the database querying language can enter the query and search in the database. The users face a huge problem as they may not be aware of the

database dependent languages. As to provide a interface to the users such that they can enter the query in the English as most of the users of the system are familiar with the English language. The users enter his query in the general English language the system is responsible for understanding the query parse and translate into an SQL query.

LUNAR (1973)

This system comes in early seventies (1973).[2] The system LUNAR science Natural language information system which was used to serve queries regarding MOON ROCKS. It syntactically analyzed language queries and then semantic analysis on the resulting parse tree. This system uses Augmented Transaction Network. Parses wood's procedural semantics.

ENGLISH WIZARD

English Wizard[1] is another successful natural language query tool for relational database. It is one of the leading software products that translate ordinary English database requests into SQL, and then return the results to the client. English Wizard enables most database reporting tools and client/server applications to understand everyday English requests for information, and it also provides graphical UI for users use.

LADDER (1978)

It was designed as a natural language interface to a database of information about US Navy ships. It uses semantic Grammar to parse questions to query a distributed data base. The LADDER. System is based on a three layered architecture. The question answering is done via parsing the input and mapping the parse tree to a database query the first component of the system is for Informal Natural Language Access to Navy Data (INLAND), which accepts questions in a natural language and produces a query to the database Second Component (IDA) would compose an answer that is relevant to the user's original quer y in addition to planning the correct sequence of file queries The third component of the LADDER system is for File Access Manager (FAM). The task of FAM is to find the location of the generic files and manage the access to them in the distributed database.

CHAT-80 (1980)

This is one of the best-known NLIDBs of the early eighties. CHAT-80 was developed in Prolog language. In this system English questions transferred into Prolog expressions (logical query language), which were evaluated against the Prolog database. Which translated the query in logical query language (LQL). The database of CHAT-80 consists of facts (*i.e.* oceans, major seas, major rivers and major cities) about 150 of the countries world and a small set of English language vocabulary that are enough for querying the database.

TEAM (1978)

A large part of the research of that time was devoted to portability issues. TEAM was designed to be easily configurable by database administrators with no knowledge of NLIDBs. TEAM includes extracting of the primary key and foreign key, it hasn't analyzed the relatio nship among entities and attributes, therefore it biased on extracting the general grammar of the words in a natural language category, thus increasing the complexity of extracting and leading to the complexity of language processing.

ASK (1996)

Ask (originally known as Ask Jeeves) is a question answering focused web search engine founded in 1996 by Garrett Gruener and David Warthen in Berkeley, California. The original software was implemented by Gary Chevsky from his own design. Warthen, Chevsky, Justin Grant, and others built the early AskJeeves.com website around that core engine. Three venture capital firms, Highland Capital Partners, Institutional Venture Partners, and The RODA Group wereearly investors.

1.1 NLDBI Architecture: A probabilistic context free grammar (or PCFG) is the simplest statistical model to analyze natural language. Usually natural language sentences are transformed into a tree structure through PCFG, and the grammar tree is analyzed according to user's requirements. CFG has a formalization capability in describing most sentence structures and so well formed that efficient sentence parser.

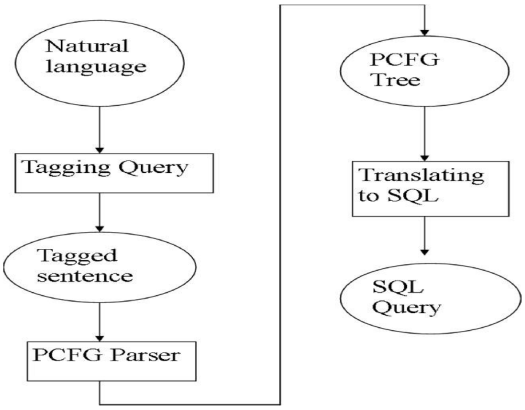


Figure 1. NLDBI Architecture

To process a query, the first step is part of speech tagging; after this step each word of the query is tagged. The second step is parsing the tagged sentence by a PCFG. The PCFG parser analyzes the query sentence according to the tag of each word, and calculates the probability of all possible grammar trees. The result of the analysis forms a grammar tree with the maximum probability. Finally, the SQL translator processes the grammar tree to obtain the SQL query by a series of dependency rules.

2. Probabilistic Context Free Grammar

Context-free grammar (CFG) and context-free grammar parsing are consistently defined concepts in the parsing literature. The trees of nested labeled constituents produced by context-free parsers, simply phrase structure trees. A probabilistic context free grammar is a CFG that assigns to each production rule a probability A phrase structure tree derivable by the grammar is defined to have a probability equal to the product of all of the production rules in the tree's derivation. A probabilistic context free grammar consists of the following: A terminal set: {wk}, where wk is a word, corresponding to a leaf in the grammar tree.

A non-terminal set: {Ni}, Ni , which is a sign used to generate terminals, corresponding to a non-leaf node in the grammar tree.

Consider a sentence w1m that is a sequence of words w1 w2 w3……wm (ignoring punctuations), and each str

in the sequence stands for a word in the sentence. The grammar tree of w1m can be generated by a set of pre-defined grammar rules. Inside and outside probability have been introduced to calculate and select the most probabilistic grammar trees. As shown in figure 3, inside probability a(p,q) is the total probability of generating words wp….. wq given that one is starting off with the non-terminal Nj. The outside probability á(p,q) is the total probability of beginning with the start symbol N and generating the non terminal Nj and all the words outside wp…. wq given the start s ymbol N.

The inside probability is defined as:

$$\beta_j\ (p, q) = p(w_{p,q} \mid N^j)$$

The outside probability is defined as:

$$\alpha_j\ (p, q) = p(w_{1(p-1)}, N^j_{pq}\ , w_{(q+1)m} \mid G\ )$$

3. SQL Translator

Translating the leaves of the tree to the corresponding SQL. The process is collecting information from the parsed tree Two techniques may be used to collect the information:

i) Dependency structure and

ii) Verb sub categorization

ing wi

i) Dependency structure A typed dependency represents Grammatical relations between individual words with dependency labels, such as subject or indirect object.

ii) Verb sub categorization If we know the sub categorization frame of the verb, we can find the objects of this verb easily, and the target of the query can be found easily.

4. Experimental Methodology

The system generated the parsed tree shown in Figure 3. using the Stanford Lexicalized Parser v1.6 that, in turn, uses the Penn Treebank.[1]

For example, for the query “Select the airlines who se seats number is more than 60 and whose mid-stops contain Wuhan.” The SQL translator scanned the parsed tree and recognized the phrases which may be the targets of the query.

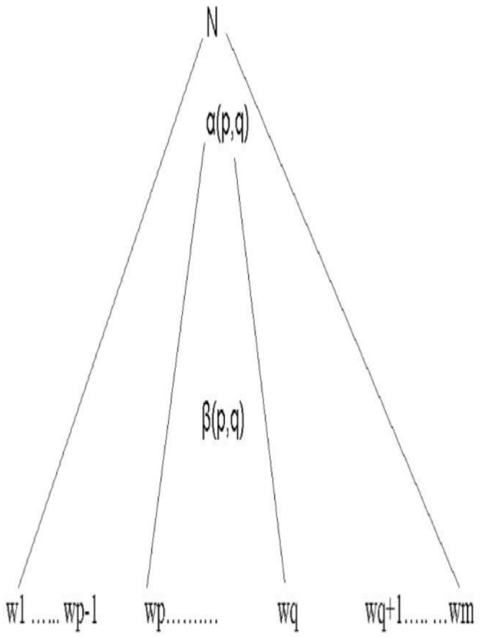
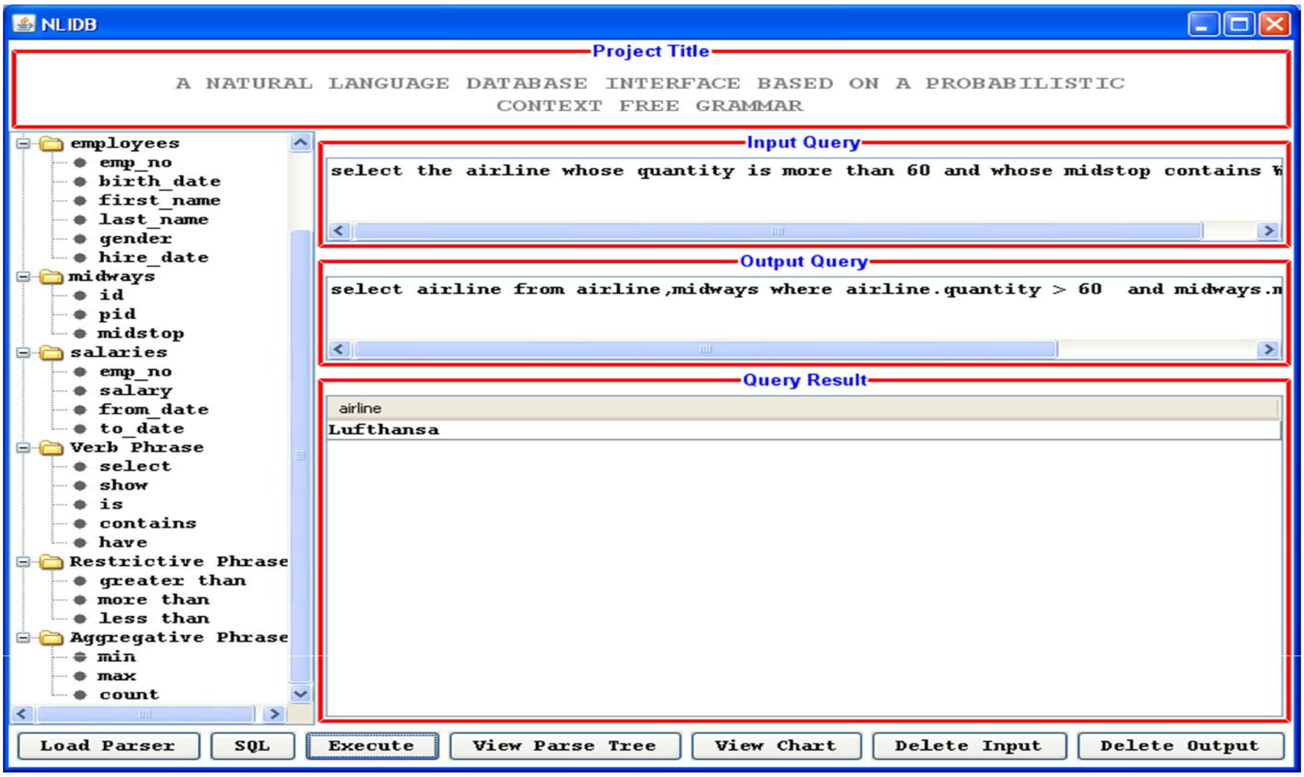


Figure 2. Query Result of input query



Project Title

A NATURAL LANGUAGE DATABASE INTERFACE BASED ON A PROBABILISTIC CONTEXT FREE GRAMMAR

employees
- emp_no
- birth_date
- first_name
- last_name
- gender
- hire_date

midways
- id
- pid
- midstop

salaries
- emp_no
- salary
- from_date
- to_date

Verb Phrase
- select
- show
- is
- contains
- have

Restrictive Phrase
- greater than
- more than
- less than

Aggregative Phrase
- min
- max
- count

Input Query

```
select the airline whose quantity is more than 60 and whose midstop contains W
```

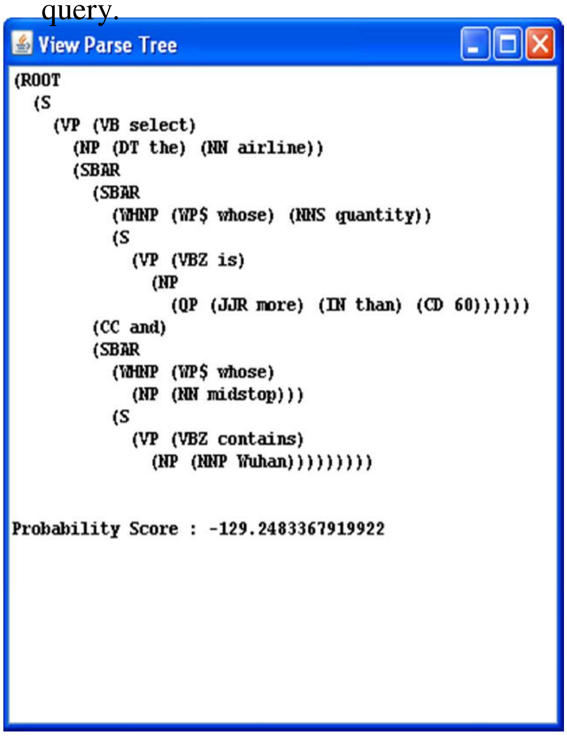
Output Query

```
select airline from airline,midways where airline.quantity > 60  and midways.m
```

Query Result

airline
Lufthansa

The first sub-tree is a verb phrase which tells the system the nearest noun phrase (the airline) is the target of the query, and this is determined by the sub-categorization frame of the verb "select". The other two sub-trees which are in the form of verb phrase indicate the modifiers of the target "airline", and they are translated into the conditions of the SQL query.



```
(ROOT
  (S
    (VP (VB select)
      (NP (DT the) (NN airline))
      (SBAR
        (SBAR
          (WHNP (WP$ whose) (NNS quantity))
          (S
            (VP (VBZ is)
              (NP
                (QP (JJR more) (IN than) (CD 60))))))
        (CC and)
        (SBAR
          (WHNP (WP$ whose)
            (NP (NN midstop)))
          (S
            (VP (VBZ contains)
              (NP (NNP Wuhan)))))))))


Probability Score : -129.2483367919922
```

Figure 3. Parse Tree of the input

Conclusions

Processing user natural language into a technical form so as to access the data from higher end data storage. NLDBI is a system that allows users to access a database in natural language and has been a popular field of study. PCFGs allow for probabilistic resolution of ambiguities. From PCFG we can calculate the probability of the input and find out the accuracy of that. The next step of is to optimize the PCFG, to accommodate more complex queries.

References

[1] Bei-Bei Huang, Guigang Zhang, Phillip C-Y Sheu "A Natural Language Database Interface Based On A Probabilistic Context Free Grammar" Wuhan University, University of California. IEEE International Workshop on Semantic Computing and Systems 2008 Page(s): 155 – 162

[2] Mrs. Neelu Nihalani , Dr. Sanjay Silakari , Dr. Mahesh Motwani " Natural language Interface for Database: A Brief review" IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011.

[3]F.Siasar1,M.Norouzifard1,S.H.Davarpanah2, M.H.Shenassa3, "Using Natural Language Processing In Order To Create Sql Queries." University of Science and Culture, Iran. International Conference on Computer and Communication Engineering 2008May 13-15, 2008 Page(s): 600 -604

[4] Xu Yiqiu Wang Liwei Yan Shi, "The Study On Natural Language Interface Of Relational Databases." Teaching And Research Section Of Educational Technology And Network

Information Center Mudanjiang Medical University. 2nd International Conference on

Environmental Science and Information Application Technology, (Volume:2) 17-18 July 2010 Page(s): 596 – 599

[5] Majdi Owda, Zuhair Bandar, Keeley Crockeet, "Conversation Based Natural Language Interface To Relational Database." Manchester Metropolitan University, International Conferences on Web Intelligence and Intelligent Agent Technology Workshops, 2007. Page(s): 363 – 367

[6] Belen, Juan Gonzalez, "Natural Language Queries In CBR System." Univ. Complutense de Madrid, Madrid 19th International Conference on Tools with Artificial Intelligence, 2007. (Volume:2) Page(s): 468 – 472

[7] Dan Klein, Christopher D. Manning. "A* Parsing: Fast Exact Viterbi Parse Selection" Stanford University Stanford, CA 94305-9040

[8] Yoshimasa Tsuruoka, Jun'ichi Tsujii. "Iterative CKY parsing for Probabilistic Context-Free Grammars." Department of Computer Science, University of Tokyo.

[9] M. Marcus, Beatrice Santorini and M.A. Marcinkiewicz: Building a large annotated corpus of English: The Penn Treebank. InComputational Linguistics, volume 19, number 2, pp313-330

[10] Androutsopoulos, I., Richie, G.D., Thanisch, P.“Natural Language Interface to Database – An Introduction”. Journal of Natural Language Engineering, Cambridge University Press. 1(1),29-81,1995

[11] Linguistic Technology. English Wizard – Dictionary Administrator's Guide. Linguistic Technology Corp., Littleton, MA, USA, 1997

[12] Johnson Mark. PCFG Models of Linguistic Tree Representations. 24(4): 613-631, 1998

[13] Laxmaiah E, Joshi Sripad "Survey Of Natural Language Interface To Databases."

[14] Matthew Crocker "Probabilistic Context-free Grammars: Inside and Outside Probabilities, and Viterbi Parsing." July 8, 2003