ISSN 2278-3083
Volume 3, No.2, March - April 2014
International Journal of Science and Applied Information Technology
Available Online at http://warse.org/pdfs/2014/ijsait01322014.pdf

A Proposed Natural Language Query Processing System

Akshay G. Satav, Archana B. Ausekar, Radhika M. Bihani, Mr Abid Shaikh
Department of Computer Engineering,ParvatibaiGenbaMoze College of Engineering
Pune-412-207
agsatav@gmail.com
abausekar1@gmail.com
radhika.bihani@gmail.com

Abstract: *As today's world is moving from offline to online all the process now are been done with the help of computer's, any data we require is present on the internet so it's a challenging task to develop a system that will provide search interface/NLP System for users without knowing any specific syntax or knowledge of a database language. Hence we present a system that will provide the search interface for users especially for online applications, search engines and many other different databases, where accuracy and efficiency are most important terms required. Various analyses shows user is not restricted to formulate any kind of query so this system provides result to users any type of query he fires to the system accurately and efficiently, even if any user make spelling mistake the system will autocorrect the spelling and experimental result show spell checking technique is efficient than Microsoft spell checker.*

Keywords-Query Mapping, Spelling Error Correction.

1. INTRODUCTION

Natural Language Processing is becoming more important in the field of Human Computer Interaction. This paper addresses a search interface to convert Natural Language Query (example English Language) into the database system understandable language without having knowledge of system language. The main goal of this system is to provide communication between human and computer without recalling any sort of database DDL or DML query syntax. The more general goal of such search interface or NLP system is to make the computer able to understand the natural language so that user can address the system as if they are addressing some other person and get the expected result they want.

The goal for the interface system are, first, to provide the spelling correction for mistakes made by the user while firing query, and second to map the natural language query into database query language . The spelling correction uses two techniques, first, using dictionary and second, without using dictionary. Since efficiency is the vital term in our system and the dictionary size is so large so we use Word pair mining technique [9] for correction of spelling mistakes in the query.

This paper proposes a system which gives interface between user and computer, in the form of database query language. Also the spelling corrections of misspelled words in query are getting correct.

Word Pair Mining Technique

Search session consist of set of user queries fired by a single user within a short time period [9]. Many of the search session consisted of misspelled word and there corrected spelling. We segment the query stream from user into sessions. If the time between the two queries exceeds a certain period of time then we put session boundary between the two queries. The time period between the two queries is kept short, the reason behind this that it is observed that user firing query is corrects the misspelled word immediately after recognising the mistake. Finally we make pairs misspelled and corrected word across the whole sessions and give it a frequency to each word pair and later on discard the word pair that has lower frequency. The following table shows some example of correct and misspelled word.

Misspelled Word	Correct Word
aacoustic	acoustic
Finlad	Finland
newpape	Newspaper
Olimpick	Olympic

Table 1 Word Pairs

Query Mapping

In this the natural language query is taken in English Language, any form of statement (WH type questions [7], word, any type of statement etc.). Then the query in English language mapped according to syntax of SQL query that provides user the accurate data from database after execution of mapped SQL query. The accuracy in mapped Query is focused here.

2. LITERATURE STUDY

The very first NLP system for database system is as old as any of the NLP research. Some of the related systems till date are given below that provided database interface for users:

Lunar

Woods in the year 1972 [1] developed a system that provided a search interface for the database system that stored information about the rock samples that were brought

from the moon for research. This system used two databases that were, first, chemical analyses and second, literature references. This system used Augmented Transit Network (ATN) parser and some Wood's semantics. This system was informally demonstrated in 1971 at Second annual Lunar Science conference.

Lifer/Ladder System

LIFER/LADDER system[2] was one of the 1st good search interface technique (i.e. NLP system)and was described in the paper by Hendrix in 1978, which used a semantic grammar for parsing the query and fired that query on a distributed database system.This system was basically developed for providing search interface to the database that stored information about the U.S navy ships. This system only single table queries and incase for multiple table it supported simple join queries with easy join conditions.

Natural Language Query Processing Using Semantic Grammar

Mrs. Gauri Rao proposed that provide search interface[7] and reduced the part of user for recalling the tedious syntax of databases, this system provided the user to get the database information in his/her language. The drawback of this system was that user has to fire queries in WH type question, it also did not supported single word query, ambiguous words care was taken while processing query. A limited data dictionary was used in which words were updated after regular period of time. All the name in the Natural language query had to be in double quotes (""). The system reduced the part of user but not to that extent that user required again he had to keep some of the things in mind that the name should in double quotes, query should consist of words that were present in the data dictionary, and the query had to be in WH question type and so on.

After doing the above literature survey we came to the conclusion that the existing systems were not that much crucial to as per the user's expectations. It reduced the user part of recognizing or remembering the syntax but not to that extent. Again the system till yet provided search interface for only limited database because of that user was forced to write the queries according to that database only, and now a day as the advancement is going on the size of database is also increasing day by day. Any of the systems proposed till today are not able to correct the mistakes of the spelling done by users, and normal human tendency is to do mistake. None of the system focused on the accuracy and efficiency that are the important factors for online applications. The most important thing was that user had to select the table in advance in our system we will get the table automatically according to the columns in the query after tokenizing. Cause we will be getting the columns previously before getting tokens.

So after analyzing all these system we have tried to develop a system that will be useful for any online application cause of its accuracy and efficiency. Our system does not forces or restricts the user imagination of firing query, he can fire any of the query he want's weather it is WH type, single word or of any type. We also provided spell checking technique so the user's part is much more reduced due to this feature. In our system we are not only focusing on the accuracy and efficiency but also on the ability of the system of correcting users mistakes. Moreover our system is not there for limited data, any domain specific data, but for unlimited data irrespective of any domain. The query generated by our system after fetching it to the database will provide the result that user will be expecting. In the next section we will have the look on the proposed system.

3. PROPOSED WORK

System Description

In this section we will be discussing the system in detail: The system we proposed in this paper converts the natural language query into SQL (Structured Query Language) which is a database programming language. We will perform the following steps for the transformation of query from natural language to database query (SQL) sequentially as listed in the following points:

- First we will accept the string in natural language.
- After accepting the query we will check the query for misspelled words (if any) using word pair mining.
- After that we will split query into tokens.
- After getting tokens we perform the SQL mapping for transformation.

The above process will be as shown in following figure:

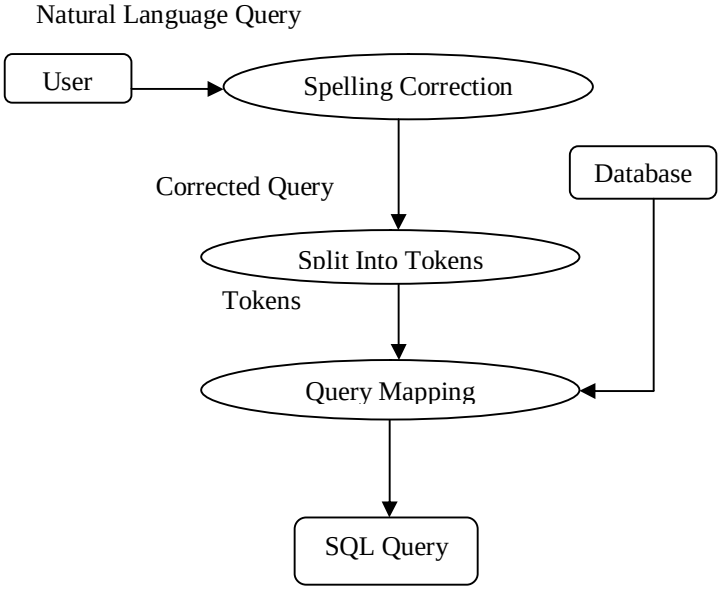

Figure 1 System Architecture

Spelling Correction

In the above process the spell correction of misspelled word[5] [6] [9] made by the user while entering the queries

is done with the help of word pair mining technique that is performed as follows:
1.The queries entered by the user will have same number of word
2.The difference will be of just a single word in the above two queries.
3.For the word pair the word in the first query will be misspelled and the second will be considered as corrected give frequency to each word pair.
4.After this we will be discarding the low frequency word pairs.
The word pair in the above technique will mine from search session at Bing.

Query Mapping
After getting the correct query we will tokenise the query, after getting the tokens we will extract the meaningful tokens from the query that will be required for the mapping or transforming the query in database query. The meaningful tokens resembles to the token that match with any of the columns, values, any aggregate function etc. that is entered by user. After extracting the meaningful tokens we than map the NLQ to database query according to our algorithm the following example will make much more clarification about the meaningful tokens
E.g.:

User Query- name of student having highest marks

Meaningful Tokens-name (column), marks (column), highest (aggregate function max ())

After getting the meaningful tokens the query will be transformed in SQL query by putting all these tokens in the SQL syntax. It is as shown in the following example:

After getting the meaningful tokens

SQL query- SELECT name, marks FROM data WHERE marks=max of (marks);

Even if the user makes some grammatical errors in query then the system will provide the grammatically correct suggestions [5] [6].

4. FUTURE SCOPE

In the above section we discussed about the proposed system for Natural Language Processing. In this section we will discuss the future scope of the system. As discussed proposed system converts the Natural Language query in SQL query, also it corrects the spelling automatically. In future we will be focusing towards reformulation of query and transforming the Natural Language Query in the SPARQL database language required for the semantic search to provide the more accurate result as the user require instead of checking the multiple link as we do at present. Today's generation also use lots of abbreviation so keeping it in mind we will be reformulating the query (for e.g. LA by Los Angeles and TOI by Times Of INDIA) so in the future we will try to implement the above discussed goals.

5. CONCLUSION

In this paper we proposed the search interface that will be applicable for the online applications, provides ease to the user by reducing their part of recalling complex database language syntax. Our system also corrects the spelling mistakes did by the users, automatically and also takes care of grammatical errors. Proposed system generates output query irrespective of the database.

REFERENCES

[1] Woods, W.A. et al, "**The Lunar Sciences Natural Language Information System**", BB&N Report 2378, June **1972.**

[2]Hendrix, G.G., Sacerdoti, E.D., Sagalowicz, D., Slocum, J. "**Developing a natural language interface to complex data**", in ACM Transactions on database systems, 3(2), pp. 105-147, **1978**.

[3]Tomek S., Fang L., Jose Perez-Carballo and Jin W. "**Building Effective Queries in Natural Language Information Retrieval**" GE Corporate Research & DevelopmentResearch Circle, Niskayuna, NY 12309

[4] JOSEPH, S.W, ALELIUNAS, R. "**A Knowledge-Based Sub System For A Natural Language Interface To A Database That Predicts And Explains Query Failures**", IN IEEE CN, PP. 80-87, **1991**.

[5]Qing C., Mu L., Ming Z."**Improving Query Spelling Correction Using Web Search Results**",Natural Language Processing and Computational Natural Language Learning, pp. 181–189, Prague, June **2007**.

[6]Huangi, Guiang Z., Phillip C-Y S. "**A Natural Language database Interface based on probabilistic context free grammar**", IEEE International workshop on Semantic Computing and Systems **2008**.

[7] Mrs. Gauri R."**Natural Language Query Processing Using Semantic Grammar**". / (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, **2010**, 219-223

[8]Aminul Islam and Diana"**Correcting Different Types of Errors in Texts**",Inkpen University of Ottawa, Ottawa, Canada diana@site.uottawa.ca, **2011.**

[9]Ziqi W., GU X., Hang L., and Ming Z."**A Probabilistic Approach to String Transformation**" IEEE transactions on knowledge and data engineering VOL:PP NO:99 YEAR **2013**.