Extracting Sql Query Using Natural Language Processing

Nandhini S, B.Viruthika, Almas Saba, Suman Sangeeta Das

Abstract: ***The basic idea is to obtain SQL query from Natural Language Query efficiently. Structured query language is the most used tool for managing data in the relational database system. Most users are not familiar with SQL which is problematic for managing and retrieving data. A model has been proposed to find a solution to the problem. This helps people who are new and inexperienced in using SQL. The speech in Hindi language is taken as input and is converted into text. This is further translated into SQL query using word mapping technique. Finally, the output is obtained from the result using database.***

Keywords: Natural Language Processing, Structured Query Language, Speech to Text, Lexical Analysis, Syntactic and Semantic Analysis, Python, Data Dictionary.

I. INTRODUCTION

The man-machine gap is narrowed down by Natural Language Processing. Intelligent computers are built that can interact with the human being. NLP is a subfield of Artificial Intelligence that is used to build intelligent computers. English sentences are interpreted by the Natural Language Query Processing. It is widely used in research fields. One can ask questions in natural language and get the required results which proves to be a very convenient and easy method of data access. It involves a lot of complexities.This project helps users who do not have knowledge about the query languageseg: SQL (Structured Query Language), along with the complexities it involves. Placement cell officers who work on student database can use the technique to extract data. Speech recognition is the ability of a machine or program to recognize words and phrases which are either spoken or written in text form to a system-understandable format. The software accepts the Natural Speech. Natural speech is lay man language which is convenient and widely used across the world. Problems arise in the analysis or generation of Natural language text, such as tokenization, syntactic analysis, semantic analysis and working with dictionaries and grammars necessary for such analysis. Semantic Grammar performs this translation.
The system has Graphic User Interface and Error handling module. On GUI, there will be three options for the user-one can view all the contents of the table or can enter a query in SQL or in natural language. Query will get converted to SQL query.Many challenges like Ambiguity and mapping are faced during the process. Ambiguity means one aspect can have more than one forms and meanings. This is a challenge in the conversion of natural language query to SQL query. In this case, one word maps to more than one sense. Also, immediately preceding sentence affects the interpretation of next sentence by the compiler. This causes problems in Query generation as the sentence formation becomes difficult.
Example: Using SELECT and INSERT query at the same time can cause ambiguity problems.
Section II describes the various work done on different papers. Section III pays attention on the implementation of the proposed architecture with a rough algorithm. Section IV finally concludes about the project.
Thanks to the modern day technologies android has emerged as one of the top notch mobile operating systems which even has an easier development interface for application developers as compared to other OS. This paved the path for easier development of complex applications for developers[10].

Revised Manuscript Received on April 10 ,2019

Ms.Nandhini.S, Assistant Professor, Dept. of CSE, SRMIST-Ramapuram Campus, Chennai, India
B.Viruthika, Dept.of CSE, SRMIST-RamapuramCampus,Chennai, India
Almas Saba, Dept.of CSE,SRMIST-Ramapuram Campus, Chennai, India
Suman Sangeeta Das, Dept. of CSE,SRMIST-Ramapuram Campus, Chennai, India

II.LITERATURE SURVEY

This section reviews the different works concerned with the project. Some papers were studies and summarized as follows:

Paper[1] Anum Iftikhar, Erum Iftikhar, Muhammad Khalid Mehmood proposed a system for "Domain Specific Query Generation from Natural Language Text. It involves generation of SQL Queries using Stanford Parser. The paper revolves around ambiguity problems in NLP. Automated queries of NoSQL can be used as an application for the idea presented in the paper. It can also be used to design NL business etiquettes.

Paper[2] Prof. DebaratiGhosal, TejasWaghmare, VivekSatam, ChinmayHajirnis proposed a system for "SQL query formation using natural language processing". SQL query extraction from NLP is the idea discussed in the paper. This gives all users all possible intermediate queries.Appropriate intermediate query is selected by the user. The system then generates SQL query. This is done from the intermediate codes. System then execute the query and give output to the user.

Paper[3]PrasunKantiGhosh, SaparjaDey, SubhabrataSengupta proposed a system for "Automatic SQL

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Query Formation from Natural Language Query". This system involves conversion of Natural Language Query to SQL language.It also presents the idea of Speech to Text Recognition. Python programming language has been used.

Paper[4]"Translating Controlled Natural Language Query into SQL Query using Pattern Matching Technique " is a system proposed by Rajender Kumar, MohitDua. Query to the database is given by the user and then the retrieval is done with the help of NLP. This system uses two Analysers such as morphological analyser and word group analyser. The main purpose of this analysers is to extract the keyword from input.Finding the type of keyword is the next step. It uses pattern matching technique to carry out this task.

Paper[5] Prof. Sonal Gore, NiketChoudhary worked on "Impact of IntelliSense on the accuracy of Natural Language Interface to Database ". An interface is used to ask query to the database in one's own language. Machine generates suggestions using intelligence. Based on previously typed words, sentences are formed. Suggestions frame a complete and correct query which can be used while connecting to the database for extracting data.

The related work includes the survey of various place people and their publications as a context of Research. The following implementation is done after studying various publications and sources.

III.PROPOSED SYSTEM

The proposed model presents the idea of extracting SQL query using natural language processing for data manipulation and extraction. Python from anaconda and Jupyter notebook has been used for implementation of the model.

The system will deploy a natural language understanding component that identifies speaker intents. The variables are needed for a specific nt example. Our solution us0es the technique presented in the paper uses corpus library, and also enhances it, which we are developing which will be schema specific.

The problem we address is a subcategory of a problem which has broader aspects like natural language to machine language. SQL is known for its differentiating, high level language and close connection to the hidden data. These characteristics are utilized in our project.SQL is tool for manipulating data. To create a system which can generate a SQL query from natural language we need to make the system which can understand natural language.

The architecture of the proposed model is described below

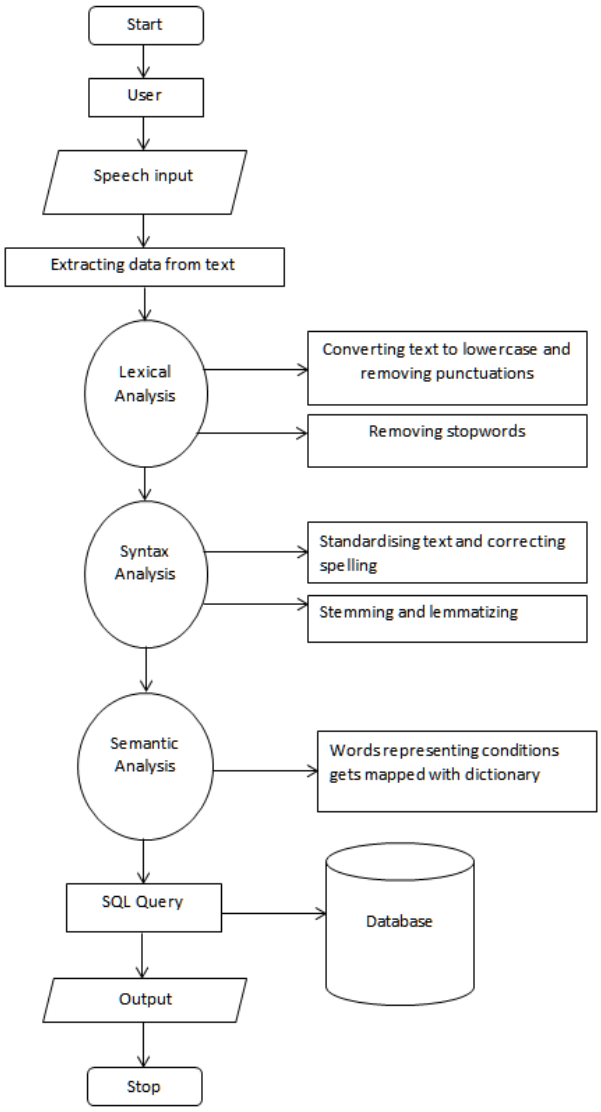

Figure 1: Architecture for the conversion of NLP to SQL

By teaching a system to identify the parts of speech of a particular word in the natural language, which is called tagging, most of the researches can be done. After this the system is made to understand the meaning of the natural query, all the words are put together which is called parsing.SQL query using proper syntax of MySQL is generated after the parsing is done.

NLP applications mostly make use of Python which is the most favoured and widely used tool. Sophisticated libraries like NLTK SpaCy, TextBlob, coreNLP are used in python.

NLTK: A suite of libraries having symbolic and statistical features.

SpaCy: A trending library having frameworks associated with deep learning.

TextBlob: A library based on NLTK and a pattern.

CoreNLP: A library that provides parts of speech and forms of base word.

The model comprises of six modules which are described below:

A.EXTRACTING THE DATA

Speech in Hindi from the user is taken as the input converted into text using PyAudio and Speech Recognition Library. The audio can be provided through a microphone or similar device. Dictate feature of Microsoft Word can also be used for the speech recognition. Later, the text from the document can be retrieved using Python.

```
r=recogniser()
with sr.Microphone() as source:
print("Please say something")
audio = r.listen(source)
print("Time over, thanks")
try:
print("I think you said:")
print(r.recognize_google(audio, language ='hi-IN'));
except: pass;
```

The resulting Hindi text would be:

ग्रीसमेंकौनसेशहरस्थितहैं

B. TRANSLATING SPEECH

The converted Hindi text is translated into English language using goslate library. Goslate library provides API for Python.It uses google translation website for conversion to English text.

```
gs = goslate.Goslate() translatedText =
gs.translate(text,'en')
```

The translated English sentence for the above input is:

Which cities are located in Greece

C. LEXICAL ANALYSIS

C.1 Converting text data to lowercase

Conversion of the text into lowercase is the primary step in lexical analysis. The lower() function converts all uppercase or capitalletters present in the string into lowercase or small letters. This makes easy to understand the sentence and convert them into tokens.

```
x="Select all the students from student table."
low=x. lower();
```

The output of which is:

"which cities are located in greece."

C.2 Removing punctuations

Punctuations like {, , :,"","',!,;} are eliminated from the sentence for tokenization. This step is very significant as punctuation does not add any extra info or value. Hence exclusion of such occurrences will have reduced the size of the data and increase computational proficiency. The replace function and regex is used for this.

```
for c in string.punctuation:
s= s.replace(c,"")
```

Output:

which cities are located in greece

C.3 Tokenizing texts

Splitting of sentences into minimal meaningful units is referred to as tokenization. Each result unit is known as tokens. It is easy to identify table name, attributes and their clauses using tokens. Libraries used in tokenization are NLTK, SpaCy, TextBlob.

['which', 'cities', 'are', 'located', 'in', 'greece']

C.4 Removing stopwords

The mutual words that convey less or no meaning associated to other keywords are called stopwords. If such words are removed more important keywords can be focused on. Examples of stopwords {'I','are','the','in','of'…}. The stopwords can be removed using NLTK library'. These libraries are pre-defined in Jupyter Lab.

```
!pip install nltk
import nltk
nltk.download()
from nltk.corpus import stopwords
stop = stopwords.words('english')
for c in list:  if c not in stop_words:
z.append(c)
```

Output:

['cities', 'located', 'greece']

D. SYNTAX ANALYSIS

D.1 Stemming

Stemming involves extraction of root words. For example, "study", "studies" and "studying" are stemmed into "study". This is also done by using the default libraries present in Jupyter Lab.

text=['I like to study', 'She studies', 'She is studying in college']

Output: ['I like to study', 'She study', 'She is study in college']

D.2 Lemmatizing

Lemmatizing is extraction of root words by checking the vocabulary. For instance: ['Got', 'secured', 'score'] is lemmatized into ['secured'].

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

E. SEMANTIC ANALYSIS

E.1 Noun-Pronoun-Verb Tagger

The extracted tokens are tagged as noun, pronoun, or verb. The various tags with their description are mentioned below

Table 1 : Parts of Speech Tags and Description

Tag	Description	Example
CC	Conjunction	And,or,but
CD	Cardinal Number	Five,3
NN	Noun	Tiger,Chair
NNS	Nouns	Cities
RB	Adverb	Extremely, Hard
VBN	Verb	Sunken

Output:

[('cities', 'NNS'), ('located', 'VBN'), ('greece', 'NN')]

E.2 Ambiguity Remover

The most apt attribute is extracted and mapped with the relation after removal of ambiguous attributes existing in multiple lines. Ambiguity means one aspect can have more than one forms and meanings.

Output: No Ambiguity

E.3 Relations-Attributes-Clauses Identifiers

Classification of relations,attributes, and clauses based on tagged elements is done after the elements into noun-verb-pronoun. This is done in order to easily recognise the relation name and its attributes to form SQL query.

Table 2: Relations-Attributes and Clauses

Token	Domain Name	Domain Type
City	City	Attribute Name-City
City_table	City_table	Relation Name-Student
Greece	Greece	WHERE Clause = Country = Greece

F. QUERY GENERATION

The final query is generated based on extracted elements after classifying the relations, attributes and clauses and further removing ambiguity. Then the query is analysed and executed, and the outcomes are extracted from the database and are displayed to naive users.

Final Output:

SELECT City FROM city_table WHERE Country="greece"

IV.CONCLUSION

In the paper, we present an idea to convert a Hindi speech into SQL query. This is first done by converting the sentence into English text, followed by formation of SQL Query, with the help of several python libraries like Goslate, NLTK. After query formation, the data is extracted from the database. The project objective is to provide accessibility to handle data without having any specific background in Computer Science. The further advancement and enhancement of the functions and keywords will be seen in future.

REFERENCES

1. "Natural Language Processing Recipes" by Akshay Kulkarni, Adarsha Shivananda.
2. "Learning Jupyter", a book by Dan Toomey.
3. W. A. Woods, R. M. Kaplan, and B. N. Webber, "The Lunar Sciences Natural Language Information System," Final Report, vol. BBN Report 2378, Bolt Beranek and Newman Inc., Cambridge, Massachusetts, 1972.
4. "The Ultimate Guide to Speech Recognition Using Python" by David Amos.
5. A. Shah, J. Pareek, H. Patel, N. Panchal, "NLKBIDB - Natural language and keyword based interface to database," International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1569-1576, 2013.
6. D. H. D. Warren, F. C. N. Pereira, "An efficient easily adaptable system for interpreting natural language queries," ACM Journal of computational linguistics, vol. 8, Issue 3-4, pp. 110-122, MIT Press Cambridge, MA, USA, Jul.-Dec. 1982.
7. G. G. Hendrix, E. D. Sacerdoti, D. Sagalowicz, and J. Slocum, "Developing a natural language interface to complex data," ACM Transactions on database systems, pp. 105- 147, 1978.
8. M. Ahmed, M. Gandhe, "Intelligent natural language query processor," Fourth IEEE Region 10 International Conference on TENCON, pp. 47-49, 1989.
9. E.F. Codd, "Seven Steps to RENDEZVOUS with The Casual User," In J. Kimbie and K. Koffeman, editors, Data Base Management. North-Holland Publishers, 1974.
10. A. Shingala, P. Virparia, "Enhancing the Relevance of Information Retrieval by Querying the Database in Natural form ," in International Conference on Intelligent systems and signal processing (ISSP), IEEE, pp. 408 - 412, 2013
11. https://arxiv.org/abs/1801.06146
12. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.
13. Huangi, Guiang Zangi, Phillip C-Y Sheu - A Natural Language database Interface based on the probabilistic context-free grammar, IEEE International Workshop on Semantic Computing and Systems 2008.
14. A Survey of Natural Language Query Builder Interface to Database, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 5 Issue 4, 2015.
15. Srirampur, S., Chandibhamar, R., Palakurthi, A., & Mamidi, R. (2014). Concepts identification of NL query in NLIDB systems. In Asian Language Processing (IALP), 2014 International Conference on (pp. 230-233). IEEE.
16. Battan, A., & Chaudhary, A. (2014).Natural Language Interface to Databases-An Implementation in International Journal of Advanced Research in Computer Science.
17. "Natural Language Understanding" book by James Allen, Pearson Education, 2002.
18. "Speech and Language Processing" book by D. Jurafsky, J.H.Martin, Pearson Education, 2002.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

AUTHORS PROFILE

Ms. Nandhini Sis an Assistant Professor in the Dept. of CSE at SRMIST-Ramapuram Campus, Chennai , she completed her M.E and is tutoring and has guided this project. Her Research interests are in the domain of Machine Learning.

B.Viruthika Dasis a B.Tech student of 3rd yr. from SRM Institute of Science and Technology, Chennai,her current interests are in the field of Android app Development,Networking and augmented reality based mobile applications

Almas Saba is a B.Tech student of 3rd yr. from SRM Institute of Science and Technology, Chennai, her current interests are in the field of Web Development and Networking.

Suman Sangeeta Dasis a B.Tech student of 3rd yr. from SRM Institute of Science and Technology, Chennai , her current interests are in the field of Python Developer and computer graphics.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication