iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VI **Month of publication:** June 2021

DOI: https://doi.org/10.22214/ijraset.2021.35804

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI June 2021- Available at www.ijraset.com

Natural Language to SQL query Generation

Kiran Raj R[1], Jithu Moni[2], Prof. Elizabeth Isaac[3], Prof. Joby George[4]
[1, 2, 3, 4]Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala

Abstract: ***Today, everyone has a personal device to access the web. Every user tries to access the knowledge that they require through internet. Most of the knowledge is within the sort of a database. A user with limited knowledge of database will have difficulty in accessing the data in the database. Hence, there's a requirement for a system that permits the users to access the knowledge within the database. The proposed method is to develop a system where the input be a natural language and receive an SQL query which is used to access the database and retrieve the information with ease. Tokenization, parts-of-speech tagging, lemmatization, parsing and mapping are the steps involved in the process. The project proposed would give a view of using of Natural Language Processing (NLP) and mapping the query in accordance with regular expression in English language to SQL.***

***Keywords: NLTK (Natural Language Toolkit), NLP (Natural Language Processing), SQLite*, Chat80**

I. INTRODUCTION

In the modern era the user has to interact with the computer for information retrieval in many areas like education, tourism, medicine etc. To receive the data from the database is quite a challenging task for a non-expert user. The user has to have a knowledge of the DBMS system in order to receive the data from the database. Management of database is done by Database Management System (DBMS). Hence a user with less knowledge faces difficulty in extracting the data. Interaction between the the computer and user is done with techniques from Natural Language Processing (NLP). Natural Language Processing is widely used for different applications like in the Tourism industry where it is used for the retrieval of places and other geographical data. The objective of this project is to generate SQL query from natural language query in order to retrieve data. The approach is to access the database using natural language input without any knowledge of Query language. The rule based approach that we propose would be much more accessible by the non-expert users. Without learning the query language the user can access the database with natural language inputs.

II. LITERATURE SURVEY

A. Title : Formation of SQL from Natural Language Query using
Author: Uma M, Sneha V, Sneha G, Bhuvana J, Bharathi B

In this fast technologically advancing world, it's became important for us humans to interact with computers to provide assistance in many fields like medicine, education, space research, etc. Accessing the data from the database is quite difficult process. Database Management System (DBMS) which is used for the manipulate the database, and a person who knows how to use it will only be able to access data. Hence a user with lack of knowledge faces difficulty in extracting the data. Natural Language Processing (NLP) Techniques are used for the user and the computer Interaction between the the computer and user is done with techniques from Natural Language Processing (NLP). Natural Language Processing is widely used for different applications like in the Tourism industry where it is used for the retrieval of places and other geographical data. Chatbots which uses textual or voice input is also an application of NLP. The aim is to simplify the data retrieval from database using SQL query generated from Natural language input.

B. Title : An algorithm to transform natural language into SQL queries for relational database
Author:Garima Singh, A. Solanki

For the efficient interaction of the database and the user there has to be an intelligent interface and that is the need for a database application. Databases must be smart enough to make the accessing of data faster. However, not every user knows SQL language or how to sue it to get the required output. Without learning the query language, the user can access the database with natural language inputs like English.

C. Title: Domain Specific Query Generation from Natural Language Text
Authors: Anum Iftikhar, Erum Iftikhar, Muhammad Khalid Mehmood

This proposed approach is to generate SQL query from NL query. The specification of software is important in the field of Natural Language Processing The slightest of mistake in the software department can affect the system. There were many issues faced while automating the process of Natural language to SQL query generation due to some software specification. The natural language texts are tested against some predefined list of sentences.

©IJRASET: All Rights are Reserved

III. TECHNOLOGIES USED

A. PYTHON

Python is a highlevel language interpretted and used for general purpose programming

B. NLTK

NLTK library for python will be used for input stemming. This library is used for computational linguistic Toolkit .

C. SQLite

SQLite is an in-process library that implements a self-contained. SQLite is an embedded database engine. The SQLite is implemented in python using the module sqlite3.

D. Char80

Chat80 module contains the functions to extract data from the World database, and can be used for the evaluation of different operation under natural language toolkit using nltk.sem.evaluate function. Chat80 is a globally accepted database.

IV. OBJECTIVE

The proposed system is to lower the communication gap between the User and the system. Improve the interaction between the accessing of database. A normal user may ot know how to to write a query to retrieve data from database as database only knows standard querie and the database schemas.The system is designed in such a way that the natural language input provided by the user will be converted into corresponding SQL query with the help of different layers residing. The system is designed to accept simple input queries given and convert it to SQL query.

V. IMPLEMENTATION

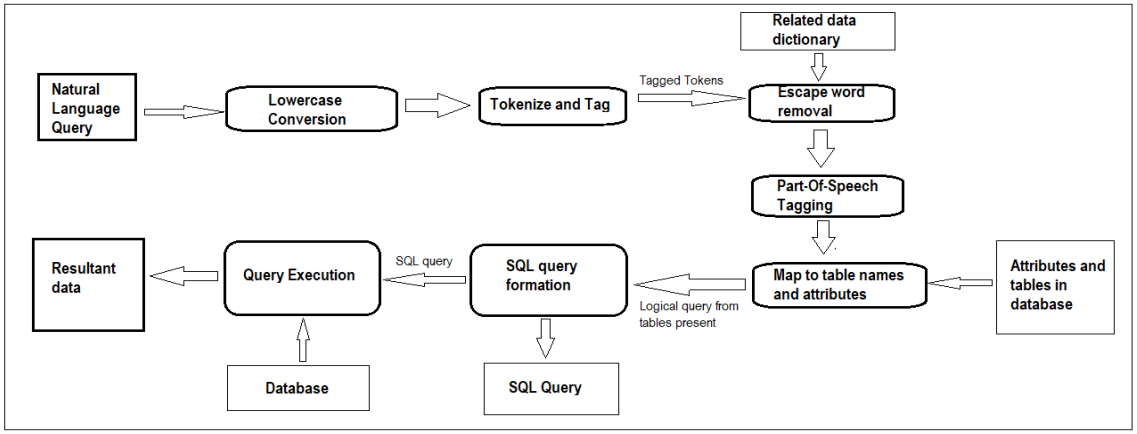


Fig. 1 Natural Language to SQL Query Implimentation Diagram

The steps in the system, is as follows:

1) *User Input*: The user input their query to the system
2) *Lowercase Conversion*: The Natural Language Query is then converted into lowercase.
3) *Tokenization*: After lowercase conversion is then converted into tokens and a token id is provided to each word of NLQ.
4) *Escape word removal*: The extra words are removed which are not needed in the analysis of query.
5) *Part-Of-Speech Tagger*: The tokens are then classified into nouns, pronouns, verb and String /integer variables.
6) *Relations-Attributes-Clauses Identifier*: Now the system classifies the tokens into relations, attributes and clauses on the basis of tagged elements and also separates the Integer and String values to form clauses.
7) *Query Formation*: After the relations, attributes and clauses are extracted, the final query is constructed.
8) *Query Execution and Data Fetching*: The query is then executed and data is fetched from the database.
9) *Results*: The final query result is displayed to the user on the User Interface.

©IJRASET: All Rights are Reserved

Direct access to the database for a common man requires quite a lot of effort and knowledge. Only queries written in the Query language is recognized by the Database management system. To make the retrieval of data more accessible to a naive user the proposed work provides facility of entering the query in English and the output SQL query will be produced only after the processing of the different modules.

The user sends a query in English text form which is then sent into several natural language processing modules. After the Natural language phase the mapping phase and where the query is identified and is used for the final SQL query generation. After the SQL query formation the data retrieval phase is left. For that the SQL query is executed with respect to the database.

Several modules are used for the keyword which is only needed for the query formation. Excess words residing would cause the decrease of performance of the system. The data goes through a series of tasks followed by a mapping phase. The NLP phase consist of separating the different tokens and lemmetization of the tokens and the tokens are tagged and in the mapping phase the attributes are identified to form the query.

VI. IMPLEMENTATION DETAILS

A. Tokenization

Tokens are the base for a Natural Language. Basic raw text processing happens at this phase. The tokens split the sentences into Word based on whitespace character. In the proposed system, we applied tokenization as soon as the text input is received from the user and the tokens obtained are stored in the form of a list. The word tokenize module of 'nltk.tokenize' library of Python is used for the process.

B. Escape Word Removal

The tokens produced after the tokenization process is analyzed here. From the tokens taken we only need the keywords that is helpful in generating or predicting the query. The extra words that are not needed, are removed.

C. Pos tagger

Here syntactic analysis is used to analysis the logical meaning of certain given sentence formed by the tokens. In this phase the exact meaning of the words are extracted. From this these tokens are then tagged into nouns, pronouns, verb, string/integer variables.

D. Relations-attribute-clause identifier

Here we use semantic analysis, we try to make sense of the tokens so that the system could proceed with the SQL query formation that is we provide meaning to the words created by the syntax. The system classifies the tokens into relations, clauses and attributes on basis of tagged elements. The process of parsing is used here. The RegExpParser() (regular expression parser) is used for parsing the POS tagged input data.

E. Query formation

The query formation is based on the structure we provide in the form of sql grammar. Using chat80 module to connect the database stored in SQLite. From the attributes, relations and clauses extracted from previous processes, following the structure given in the sql grammar and using the database connected an SQL query in constructed.

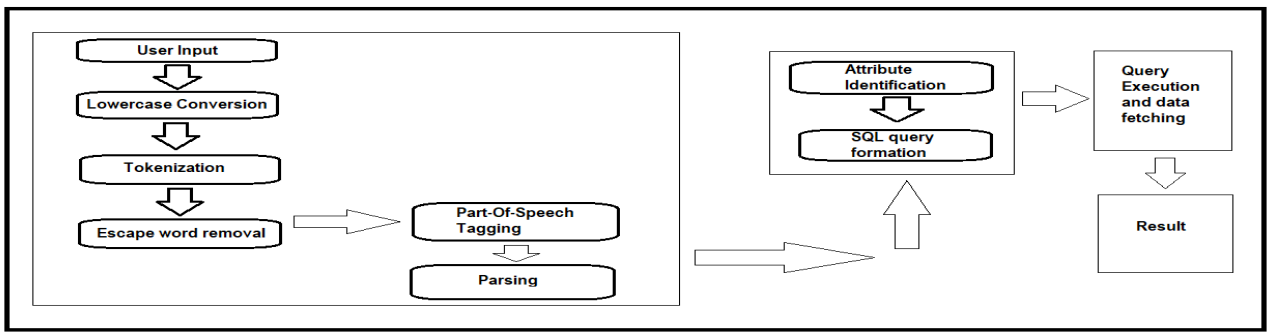


Fig.2 Workflow Diagram

©IJRASET: All Rights are Reserved

The system accepts Natural language input. The natural language query is analyzed for extra stop words and are removed which are not requires for the formation of query. Now the system seperated the required nouns and pronouns required for the query generation.Then the words which are not needed are needed are removed . After the required data are extracted the query is formed and executed.

VII. RESULT ANALYSIS

In this project natural language processing approach is used to generate the query. The query is based on the tagged tokens, the noun map is prepared. The city database is taken for reference which consist of 3 tables with *attributes City, Country, Population, Longitude, Latitude, Mapid,* we've implemented this system in Python 3.9 with NLTK module being used. The database is linked via SQLite. When we enter the input to the system it is lowercase converted and is tokenized, and removing the unnecessary words using the stopword removal process where it contains a list of words that are of no use. Them it is lemmatized and the the words are tagged. The SQL query is generated with respect to the tagged attribute that are generated.

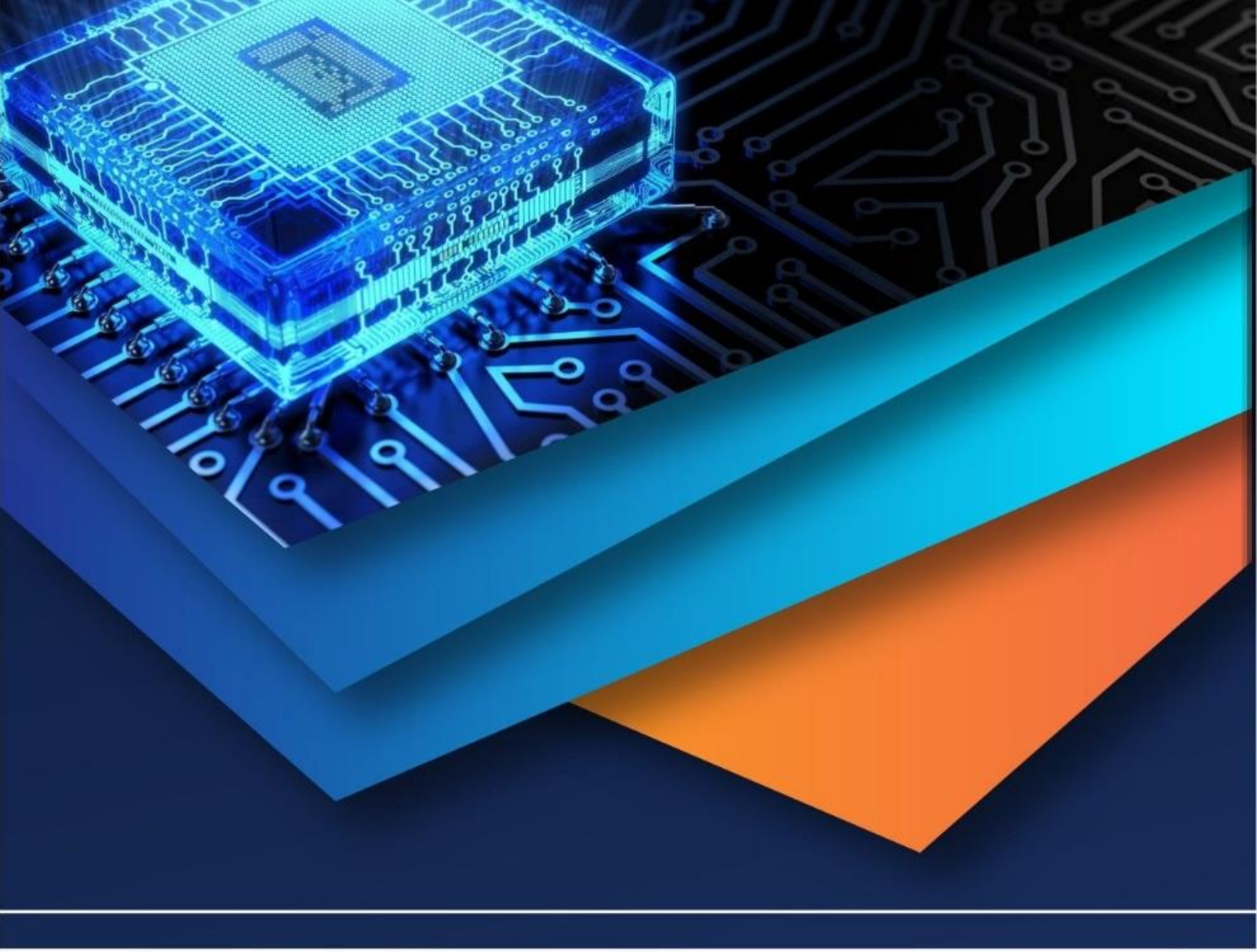
VIII. CONCLUSION

Although several methodologies are employed to extract information from a database, Natural Language Processing has set a new standard in doing the same. This work presents the steps that are involved in Natural Language Processing. The different NLTK modules are used in this project. Through this system simple queries can be generated via English input and it can be improved with more functionalities. Chatbots and for the voice interaction to the database system. There are many steps to improve the work we have done. There can be some implementations used for improving the accuracy of the outputs received and the different techniques. This can be extended to more languages and voice inputs as well. Chatbots can be created for various database access purposes.

REFERENCES

[1] A rule based approach for NLP based query processing https://ieeexplore.ieee.org/document/7391926
[2] Uma, M., Sneha, V., Sneha, G., Bhuvana, J., & Bharathi, B. (2019). Formation of SQL from Natural Language Query using NLP. 2019 International Conference on Computational Intelligence in Data Science (ICCIDS). doi:10.1109/iccids.2019.8862080
[3] P. Gupta, A. Goswami, S. Koul and K. Sartape, "IQS-intelligent querying system using natural language processing," 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), 2017, pp. 410-413, doi: 10.1109/ICECA.2017.8212846
[4] A. Iftikhar, E. Iftikhar and M. K. Mehmood, "Domain specific query generation from natural language text," 2016 Sixth International Conference on Innovative Computing Technology (INTECH), 2016, pp. 502-506, doi: 10.1109/INTECH.2016.7845105..
[5] Bhadgale, Anil M, Gavas, Sanhita R, Patil, Meghana M and Pinki, R, (2013) "Natural lan guage to SQL conversion system," International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR), Vol. 3, Issue 2, pp. 161–166
[6] Sathick, K Javubar and Jaya, A, (2015) "Natural language to SQL generation for semantic knowledge extraction in social web sources," Indian Journal of Science and Technology, vol. 8, Issue 1
[7] Natural Language Query Processing Using Semantic Grammar, Gauri Rao et al (IJCSE) International Journal on Computer Science and Engineering Vol.02,No.02, 2010, 219-223 ISSN 0975-3397

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)