ISSN No:-2456-2165

A Comprehensive Survey on Human-to-Database Communication using NLP

Sivani JC., Sathyalakshmi S., Sthuthi B., Prof. Kamleshwar Kumar Yadav
Department of Computer Science and Engineering
Global Academy of Technology

Abstract - In recent years, there is an exponential growth in the amount of data that is being generated every day. Nowadays there is the widespread use of technology in all fields. As data is growing accessing data that is required out of the huge amount of data is an important task. Structured query language (SQL) is commonly used to access data from a database. Even though these help in fetching the required data, it is not as user-friendly as using natural language. In this paper, the query writing task will be done by the model which will reduce the burden of a user who does not have any prior knowledge about the query language. The model is built using natural language processing and the deep learning model LSTM (Long Short-Term Memory).

Keywords:- *Natural Language Processing, SQL, Lexical Analysis, Syntactic and Semantic Analysis, Partial search, Long Short-Term Memory, Natural language query, Structured query language.*

I. INTRODUCTION

Currently, technology-based information retrieval is widely used in fields such as academia, education, and companies with enormous amounts of data. To retrieve this data, users must be proficient in the database language SQL (Structured Query Language) to form queries. Therefore, the effectiveness of data retrieval depends on the user's knowledge of SQL. Even if a user does not have prior knowledge of SQL, they may still need to learn it to access the necessary data from a database. Whereas NLP (Natural Language Processing) helps in formation of queries helping the users who do not know about SQL. Natural Language Processing (NLP) is a subfield of Artificial Intelligence, and its main goal is to build intelligent computers that can interact and understand human language, and that it is used to interpret English sentences by computers. Natural language query (NLQ) systems are becoming increasingly popular because they allow users to access data in a more intuitive and user-friendly way. These systems use natural language processing (NLP) techniques to understand and interpret the queries written in natural language, and then translate them into structured queries that can be executed against a database.

There are several benefits in using NLQ systems. First, they can reduce the learning curve for users who are not familiar with the query language. This can make it easier for non-technical users, such as business analysts or end-users, to access and analyze data. Second, NLQ systems can improve the efficiency of data access and analysis by allowing users to express their queries in a more concise and natural way. This can save time and reduce the risk of errors that can occur when writing complex queries in a structured language. There are also some challenges to building and deploying NLQ systems. One of the main challenges is ensuring that the system can accurately understand and interpret a wide range of natural language queries. This requires robust NLP capabilities and a deep understanding of the underlying data and its structure. Additionally, NLQ systems may need to be integrated with other tools and systems, such as visualization or reporting tools, to provide a complete end-to-end solution for data access and analysis.

In this project, the system will accept user's query in natural language as an input. The model will check whether the query is valid or not. Then it will generate tokens by performing tokenization. Each token represents a single word in the user's query. The tokens from the query clause is compared with clauses already stored in the dictionary then the algorithm scans the tokens and tries to find attributes present in the query. Then we find all the tables in the database which contain the attributes by comparing syntax and semantics. Then we build the final SQL query and execute it on the database and return the result dataset to the user.

II. LITERTURE SURVEY

The study's model utilized pre-processing steps and NLP techniques such as tokenization, lexical analysis, and semantic analysis. It employed a dictionary to link tokens with attributes and provided the necessary outcomes. However, the model has limitations and does not support all types of SQL queries. Future improvements could be made by implementing various algorithms to enhance the system's capabilities.[1]

In this study, the model utilized various pre-processing steps, such as morphological, lexical, syntactic, and semantic analysis, to prepare natural language queries for the model to understand and process easily. Once the pre-processing is completed, an unsupervised learning algorithm called GLoVe is employed to perform word embedding on the input sentence, converting it into a numerical format for more effective analysis by the model. The model has shown high efficiency in generating SQL queries with a 60% accuracy rate in matching exact queries, indicating potential for natural language processing tasks, particularly for simpler queries. [2]

The model utilized several pre-processing methods such as tokenization, lemmatization, and part-of-speech tagging to analyze the input data. Moreover, the model implemented algorithms for table linking and natural joining. To evaluate the model's performance, 50 single

ISSN No:-2456-2165

sentence natural language input queries were tested, resulting in an accuracy rate of around 82%.To improve the efficiency of the system, one possible solution is to incorporate machine learning algorithms to determine the most efficient SQL query from all the possible queries for a user's query.[3]

To process natural language queries, the model in this study employed several steps, such as morphology, lexical analysis, syntactic analysis, and speech-to-text recognition. The model was able to handle both simple and some complex queries effectively. However, it is essential to note that the system does not currently support some SQL query types, and further development is required to enhance the model's capabilities in this area.[4]

The model in this study goes through several stages of processing, including data extraction, speech translation, lexical analysis, semantic analysis, and query generation. Its purpose is to convert Hindi speech into English text and generate SQL queries based on the converted text. However, the model's capabilities are currently limited in handling complex queries and may be more effective for simpler queries.[5]

The study discussed the use of Recurrent Neural Networks (RNN) with Long Short-term memory (LSTM) and advanced encoding techniques to convert human language utterances into vectorized representations, which can be translated into logical forms in the form of sequences or trees. The evaluation metrics used in this study were exact matching and SQL Hardness Criteria. The results showed that the model performed well in handling different types of queries. However, there is still room for improvement, particularly in handling complex or challenging queries. The study highlights the potential of advanced machine learning techniques in natural language processing and its application in database management systems.[6]

III. SYSTEM ARCHITECTURE

Our proposed system is a natural language processing (NLP) tool that can help users query relational databases using natural language rather than SQL. One of the key challenges in building such a system is to accurately understand and interpret the user's natural language input and translate it into a valid SQL query that the database can understand and execute. This requires the system to have a strong understanding of both natural language and the structure and schema of the database. Our system is designed to handle ambiguities in natural language queries (NLQ) by relying on database and semantic information about the tables and attributes in the database. This can help the system disambiguate attributes with the same name and generate more accurate queries.

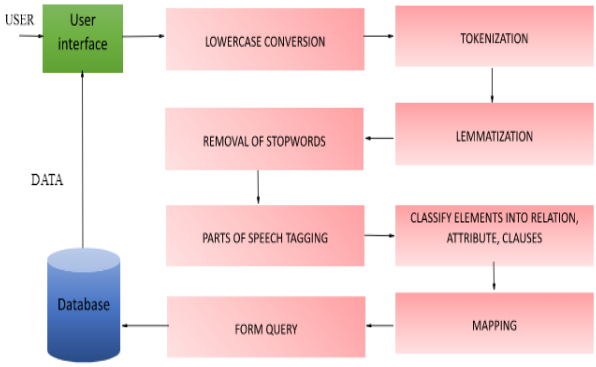

Fig 3.1. Architechture of our system

It is also important that the system can handle different relational database schemas, as this will allow it to be used with a wider range of databases. By being able to perform flexible queries and intelligent information processing, the system can adapt to different schema structures and provide more useful results to the user.

Overall, our proposed system has the potential to be a useful tool for improving the communication between humans and computers and making it easier for users to query databases using natural language.

IV. EXISTING SYSTEM

There have been several existing solutions for the conversion of human language into SQL queries using NLP techniques. Some of the most popular approaches include:

- **Rule-based Systems:** Rule-based systems use predefined grammar rules and templates to generate SQL queries from natural language inputs. While these systems are relatively simple to implement, they can be limited by the complexity of the SQL language and the variability of human language.
- **Statistical Machine Learning:** Statistical machine learning methods, such as linear regression and decision trees, can be used to learn the mapping between natural language inputs and SQL queries. These methods require large, annotated datasets to be effective, and can struggle with the ambiguity and variability of human language.
- **Deep Learning:** Deep learning methods, such as recurrent neural networks (RNNs) and transformer models, have shown promising results in the conversion of human language into SQL queries. These methods are capable of learning complex relationships between natural language inputs and SQL queries and can handle variability and ambiguity in human language.
- **Hybrid Approaches:** Some researchers have proposed hybrid approaches that combine the strengths of rule-based systems, machine learning, and deep learning to improve the accuracy and reliability of NLP-based systems for conversion of human language into SQL.
- **Attention-based Models:** Attention-based models, such as the Transformer architecture, have been used to generate SQL queries from natural language inputs. These models use an attention mechanism to focus on the most relevant parts of the input and generate more accurate and concise SQL queries.

- **Reinforcement Learning:** Reinforcement learning has the potential to greatly improve the accuracy of NLP-based systems for conversion of human language into SQL by allowing the system to learn from its mistakes and adjust over time.

In conclusion, there are several proposed solutions for the conversion of human language into SQL queries using NLP, each with its own strengths and limitations. Researchers continue to explore new methods and approaches to improve the accuracy and reliability of these systems, and it is an active area of research.

A. Attention-based Models

Attention-based models are a popular solution for the conversion of human language into SQL queries. These models use an attention mechanism to focus on the most relevant parts of the input and generate more accurate and concise SQL queries.

Attention-based models typically consist of two main components: an encoder that processes the natural language input, and a decoder that generates the SQL query. The attention mechanism is used to weigh the importance of different parts of the input, allowing the model to focus on the most relevant information and generate more accurate SQL queries.

These have shown promising results in several NLP tasks, including the conversion of human language into SQL queries. These models can handle the variability and ambiguity of human language and can generate complex SQL queries that accurately reflect the intent of the user.

Additionally, attention-based models are flexible and can be fine-tuned to perform well on specific domains or datasets. This allows them to be easily adapted to new applications and requirements, making them a versatile solution for the conversion of human language into SQL queries.

Overall, attention-based models are a promising solution for the conversion of human language into SQL queries, and they have the potential to greatly improve the accuracy and reliability of NLP-based systems in this domain.

B. Reinforcement learning

Reinforcement learning (RL) is another promising solution for the conversion of human language into SQL queries. In RL, an agent learns how to perform a task through trial and error, receiving rewards for good performance and penalties for poor performance. This allows the agent to improve its performance over time.

Here's an outline of a simple RL-based solution for the conversion of human language into SQL queries:

- **Define the task:** The first step is to define the task that the agent will perform. In this case, the task is to generate an SQL query from a natural language input.
- **Define the state space:** The state space consists of the information that the agent uses to make decisions. For this task, the state space might include the natural language input, the current SQL query generated by the agent, and any feedback received from the database.
- **Define the action space:** The action space consists of the possible actions that the agent can take. In this case, the action space might include adding, deleting, or modifying a clause in the current SQL query.
- **Define the reward function:** The reward function determines the reward that the agent receives for a given state and action. In this case, the reward might be based on the accuracy of the SQL query generated by the agent, as determined by a database.
- **Train the agent:** The agent is trained through a series of trial-and-error interactions with the environment. At each step, the agent observes the current state, selects an action, and receives a reward based on the reward function. The agent uses this feedback to update its policy, gradually improving its performance over time.

C. Deep Learning Approach

Deep learning models, such as recurrent neural networks (RNNs) and transformer networks, can be applied to the conversion of human language into SQL queries. These models can be trained on a large corpus of natural language inputs and SQL outputs.

Here's an outline of a deep learning-based solution for the conversion of human language into SQL queries:

- **Pre-process the data:** The first step is to pre-process the data. This will include converting the natural language inputs and SQL outputs into numerical representations, such as word embeddings or one-hot encodings.
- **Train the model:** The next step is to train the deep learning model on the pre-processed data. A typical approach with an encoder-decoder architecture, where the encoder processes the natural language input, and the decoder generates the SQL output.
- **Evaluate the model:** After training, the model can be evaluated on a validation set to assess its performance. Metrics such as accuracy, precision, and recall can be used to evaluate the model's ability to generate correct SQL queries.

D. Statistical Machine Learning Approach

Statistical machine learning is a method that uses statistical models to learn patterns and relationships in data. In the context of converting natural language to SQL queries, statistical machine learning can be used to train a model on a large corpus of text and its corresponding SQL queries to learn the relationship between them. The model can then be used to predict the SQL query that corresponds to a new piece of natural language text.

To apply statistical machine learning to the problem of natural language to SQL query conversion, the first step is to build a training dataset consisting of pairs of natural language text and their corresponding SQL queries. This dataset can be created manually, by mapping a set of natural language queries to their corresponding SQL queries, or through automatic data extraction from existing databases.

ISSN No:-2456-2165

Once the training dataset is created, a statistical model is trained to learn the patterns and relationships between the natural language text and its corresponding SQL queries. There are various machine learning algorithms that can be used for this purpose, including decision trees, random forests, and support vector machines.

The trained model can then be used to predict the SQL query that corresponds to a new piece of natural language text. The input text is first pre-processed to remove stop words and other irrelevant words, and then fed into the trained model, which outputs the corresponding SQL query.

Overall, statistical machine learning can be effective for more complex domains where the language used is more varied and nuanced. However, it requires a large amount of training data and may be less interpretable compared to rule-based systems.

E. Rule Based System

A rule-based system is an artificial intelligence technique that uses a set of rules to perform a specific task. In the context of converting natural language to SQL queries, a rule-based system can be used to translate sentences in natural language into corresponding SQL queries by applying a set of predefined rules. These rules may be based on linguistic and semantic analysis of the input text and can vary depending on the specific application domain and language.

A rule-based system typically consists of three main components: a parser, a knowledge base, and an inference engine. The parser analyses the input sentence and identifies the relevant parts of speech, such as nouns, verbs, and adjectives. The knowledge base contains a set of rules that describe how the input text should be translated into SQL queries. The inference engine uses the rules to generate the appropriate SQL queries based on the input text.

For example, a rule-based system for converting natural language to SQL queries could have rules such as "if the input text contains the word 'select', the following part of the sentence should be translated as the select statement in SQL", or "if the input text contains a noun followed by a verb, the following part of the sentence should be translated as the from clause in SQL". These rules can be further refined and customized based on the specific requirements of the application.

Overall, rule-based systems can be effective for simple or limited domains where the language used is relatively consistent and the rules can be easily defined. However, they may not be as effective for more complex domains or where the language used is more varied and nuanced. In these cases, machine learning-based approaches may be more appropriate.

V. ISSUES AND FUTURE TRENDS

Despite recent advancements in Natural Language Processing (NLP)-based conversion of human language into Structured Query Language (SQL) queries, there are still several unresolved issues that impede the accuracy and reliability of these systems. Natural language is inherently ambiguous and variable, posing a challenge for NLP-based systems to accurately translate it into SQL queries.

Furthermore, SQL is a complex language, and database schemas can vary widely, making it challenging to develop a generalizable solution for NLP-based conversion of human language into SQL. Additionally, the development of accurate NLP-based systems for conversion of human language into SQL requires large, annotated datasets that are currently lacking. Finally, for widespread adoption, NLP-based systems must integrate seamlessly with existing database systems.

- There is a growing trend towards the development of NLP-based systems for conversion of human language into SQL queries.
- Future trends in this field include the integration of semantic knowledge, improved handling of variability and ambiguity, the development of domain-specific systems, and the incorporation of reinforcement learning.
- NLP-based systems for conversion of human language into SQL queries are likely to become better equipped to handle the challenges of natural language as NLP technologies continue to evolve.
- The development of domain-specific systems tailored to specific domains and database schemas can address the limitations of general-purpose NLP-based systems.
- Reinforcement learning has the potential to improve the accuracy of NLP-based systems for conversion of human language into SQL and is an area of active research.

Despite challenges, the potential benefits of NLP-based systems for making database querying more accessible to a wider range of users make it an exciting area of study.

VI. CONCLUSION

There are various approaches for generating SQL queries from natural language input. As seen in the existing approaches there are different techniques through which the objective can be achieved. Some work for simple queries and some also involve solving complex queries.

NLP techniques such as lowercase conversion, tokenization, and part-of-speech tagging can help the system understand and analyse the structure and meaning of the natural language input. The relations-attributes-clauses identifier and ambiguity removal steps can then help the system disambiguate the input and generate a more accurate query.

On the other hand, deep learning algorithms such as RNNs with LSTM can be used to build more complex and efficient models that can handle a wider range of input and generate more accurate queries. These models can be trained on large datasets of natural language queries and corresponding SQL queries, allowing them to learn the patterns and relationships between the two and generate more accurate queries.

Overall, it seems that both rule-based and deep learning approaches have their own strengths and can be used effectively in generating SQL queries from natural language input, depending on the specific requirements and goals of the system.

ACKNOWLEDGMENT

We extend our sincere appreciation to Mr. Kamleshwar Kumar Yadav, who guided us throughout this research as an Assistant Professor at Global Academy of Technology. The valuable resources and support provided by the institution played a crucial role in making this research possible. We also express our gratitude to the open-source communities that provided us with the necessary tools and solutions for implementation. Moreover, we recognize the contributions of individuals who have worked towards the development and progression of NLP and SQL technologies. Their efforts have greatly facilitated our research work.

REFERENCES

[1.] Akshar Prasad, Shobha G, Deepamala N, Sourabh S Badhya," Values for Conversion of Natural Language to SQL Queries on HPCC Systems". (2019)
[2.] Dhairya Chandarana, Deepchand Dubey, Mohit Mathkar, Prof. Anagha Patil," Natural Language Sentence to SQL Query Converter",(2021)
[3.] Amit Pagrut, Ishant Pakmode, Shambhoo Kariya, Vibhavari Kamble," Automated SQL Query generator by understanding a natural language statement" (2018)
[4.] Kanti Ghosh, Prasun, Saparja , Subhabrata . "Automatic SQL Query Formation from Natural Language Query" (2014)
[5.] Nandhini S, B.Viruthika, "Extracting Sql Query Using Natural Language Processing", (April 2019).
[6.] Anisha T S, Rafeeque P C, Reena Murali , "Text to SQL Query Conversion Using Deep Learning: A Comparative Analysis". (2019)
[7.] Ikshu Bhalla and Archit Gupta, "Generating SQL queries from natural language"
[8.] Pranali Nagare And Smita Indhe, "Automatic SQL Query Formation from Natural Language Query", International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395 -0056, Volume: 04, (Mar -2017)
[9.] K. Javubar Sathick and A. Jaya, "Natural language to SQL Generation for Semantic Knowledge Extraction in Social Web Sources", Indian Journal of Science and Technology, DOI: 10.17485/ijst/2015/v8i1/54123, (January 2015)
[10.] Muhammad Khalid Mehmood And Anum Iftikhar, Erum Iftikhar, "Domain Specific Query Generation from Natural Language Text" , The Sixth International Conference On Innovative Computing Technology ,978-1-5090-2000-3/16(2016)
[11.] Tejas Waghmare And Vivek Satam, " SQL Query Formation Using Natural Language Processing (NLP)", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 3, (March 2016)