e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 10, October 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

© 2024 IJMRSET | Volume 7, Issue 10, October 2024| DOI: 10.15680/IJMRSET.2024.0710067

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Text-to-SQL Conversion by using Deep Learning/Machine Learning: Integrating Natural Language with Database Queries

Amar Kaygude[1], Onkar Rajguru[2], Sandesh Karad[3], Prof.G.T.Avhad[4]

Department of Computer Engineering, Vishwabharati Academy's College of Engineering, Ahmednagar, Maharashtra, India

ABSTRACT: Accessing and extracting insights from databases is often a challenge for non-technical users unfamiliar with Structured Query Language (SQL). This project proposes a novel solution by developing a system that converts plain language into SQL queries using cutting-edge deep learning (DL) and machine learning (ML) techniques. The system enables users, including academics, business professionals, and students, to interact with databases without needing to understand complex SQL syntax. Through a user-friendly interface, users can input queries in natural language and receive accurate, real-time database responses. Advanced DL models such as Transformer architectures (BERT, GPT, and T5) are employed to understand language structure, context, and syntax, enabling the generation of accurate SQL queries, even for complex operations involving joins, nested statements, and conditions. This approach democratizes data access, reducing the reliance on technical staff, increasing productivity, and lowering training costs. The technology encourages a data-driven culture by enabling direct interaction with databases, benefiting industries like banking, healthcare, education, and customer service. However, challenges such as maintaining high accuracy, handling ambiguous queries, and managing diverse database schemas remain. The system's robustness depends on extensive training across real-world datasets, while interpretability and error management are crucial for mission-critical applications.

KEYWORDS: Natural Language Processing, Deep Learning, Machine Learning, SQL Query Generation, Transformer Models, Data Access Automation, Database Interaction etc.

I. INTRODUCTION

This project presents a novel system that converts plain language into SQL queries, enabling non-technical users to access and extract insights from databases without needing to understand SQL syntax. Utilizing advanced deep learning models, including Transformer architectures like BERT, GPT, and T5, the system allows users such as academics and business professionals to input natural language queries and receive accurate database responses in real-time. Key features include a user-friendly interface, query previews, and autocomplete functions to enhance usability. However, challenges remain in achieving high accuracy for complex queries, handling ambiguous inputs, and adapting to diverse database schemas. Extensive training on varied datasets is essential for robustness, while interpretability and error management are critical for applications in regulated industries. This approach not only democratizes data access but also fasters a data-driven culture across sectors like banking, healthcare, and education.

1.1 Text To SQL

Relational databases are now widely used to store vast amounts of structured data from the Internet as a result of its growth. For structural data, the relational database offers easy query features together with reliable storage. While relational databases may be effectively accessed by proficient programmers using structured query languages (SQL), individuals without an understanding of SQL can access the databases with the help of natural language interfaces to databases (NLDB). As a result, the academic and industrial groups have taken an interest in text-to-SQL, which attempts to convert natural language (NL) descriptions and queries into SQL. As we cover in Chanter 3.1, the majority of text-to-SQL techniques now use neural networks, which are mostly based on the seq2seq model structure.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

1.2 Text-to-SQL Assumption

Text-to-SQL solutions assume knowledge of the underlying database structure and convert natural language (NL) descriptions into SQL queries. Instead of operating as conventional Question-Answering (QA) models, these systems are designed specifically to produce SQL queries that retrieve data from databases. Although they are able to respond to some data- related enquiries, the way enquiries are phrased has to conform to the SQL syntax and database structure. Instead of immediately rendering judgements or carrying out logical comparisons, queries must specify what data to obtain. For instance, users should query, "How old is Lily?" to obtain her age rather than, "Is Lily older than 18?" which suggests a rational judgement. After that, the user has to manually check to see whether Lily is older than 18. Likewise, one might infer Lily's age indirectly by asking, "What is her nationality if she is older than 18?" She is older than 18 if the inquiry yields nationality information. Lily must be less than eighteen if no information is supplied. This method highlights the fact that Text-to-SQL systems do not comprehend or assess the meaning of queries that go beyond the SQL syntax; they only provide users access to data. Database fields and queries must match, with the goal of obtaining pertinent data for users to examine on their own. As a result, rather than handling quality control or reasoning duties directly, the technology makes structured data access easier.

Table -1:

Name	Age	Gender	Nationality	Phone Number
Amar	21	male	indian	7744949305
Sandesh	22	male	indian	7845544489
Onkar	25	male	indian	7249454545
Lily	30	female	UK	4548455447

II. LITERATURE SURVEY

Title: OPT-IML: Instruction Meta-Learning for Zero-Shot and Few-Shot Generalization
Authors: Srinivasan Iyer et al. (2023)
Description:
This paper explores fine-tuning large language models using instruction-tuning to enhance generalization to unseen tasks. The authors introduce OPT-IML Bench, a benchmark of 2000 NLP tasks, and develop the OPT-IML models (30B and 175B), which outperform existing OPT models and compete with models specialized for specific benchmarks.

Title: NeuralDB: Querying Databases Without Pre-Defined Schemas Using Natural Language
Authors: James Thorne et al. (2020)
Description:
This paper presents NeuralDB, a database system that processes queries and updates in natural language without using predefined schemas. By using NLP transformers, the system can handle select-project-join (SPJ) queries but struggles with scaling and aggregation. To overcome this, the authors propose an architecture with multiple parallel Neural SPJ operators and an aggregation operator, achieving high accuracy on large datasets

Title: Multi-Task Learning in Natural Language Processing: An Overview
Author: Chen, S., et al. (2024).
Description:
This paper reviews the application of Multi-Task Learning (MTL) in Natural Language Processing (NLP), categorizing MTL architectures into four classes and discussing optimization techniques for effective training. It highlights the benefits of MTL in addressing overfitting and data scarcity issues while presenting benchmark datasets and potential research directions in the field.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

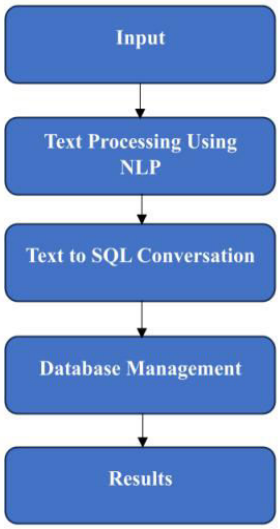
III. PROBLEM STATEMENT

Accessing and extracting insights from databases poses significant challenges for non-technical users who are unfamiliar with Structured Query Language (SQL). The complexity of SQL syntax often creates barriers to effective data interaction, limiting the ability of academics, business professionals, and students to leverage valuable information stored in databases.

IV. OBJECTIVES

Create a tool that turns everyday language into SQL queries, making it easier to get information from databases.
Create a user-friendly interface that allows individuals to interact with the system effortlessly, enabling them to access database information quickly and intuitively.
Make a tool that translates natural language into SQL to help users get answers from databases without learning complex query languages**.**

V. PROPOSED SYSTEM



Input: This is the first step in which consumers submit their questions in natural language. The design focuses usability, enabling people without technical knowledge to engage with the system easily.

Text Processing using NLP: After receiving input, the system uses Natural Language Processing (NLP) methods to evaluate and enhance the query. This includes techniques like tokenization, which divides the text into individual words or phrases; stop word removal, which removes popular terms that may not contribute value; and lemmatization, which reduces words to their simplest forms. This phase ensures that the query is clear and accurate, which improves the correctness of future actions.

Text to SQL. Conversion: After the text is progessed, the system translates the refined input into SQL queries. This entails determining the user's intent and retrieving relevant items from the query. The conversion is critical because it converts the natural language into a structured format that the database can understand, allowing for accurate data retrieval.

Database Management: The resulting SQL queries are then sent to the database management system. The queries are conducted against the database to get the necessary data. This stage is critical because it links the natural language input

© 2024 IJMRSET | Volume 7, Issue 10, October 2024| DOI: 10.15680/IJMRSET.2024.0710067

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

to the structured data contained in the database.

Results: Finally, the obtained data is given to the user in an understandable way. This might include tables, charts, or other visual aids that make the material more accessible and easier to understand. This step concludes the workflow by giving users with the information they sought, hence improving decision-making processes.

VI. EXISTING SYSTEM

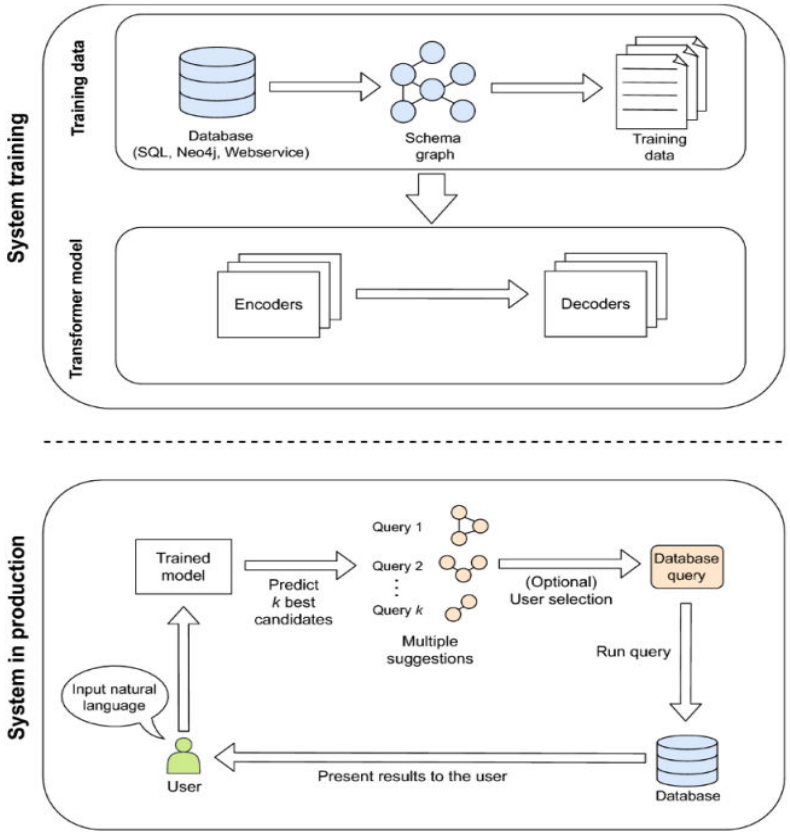


fig1:Existing System

Numerous text-to-SQL systems have been developed to convert natural language inputs into SQL instructions, simplifying database searches. Early systems relied on template-based methods, which struggled with diverse queries due to their dependence on predefined patterns. Rule-based systems improved upon this by using linguistic criteria but still faced challenges with ambiguous queries and complex database interactions. The advent of neural network models, such as Seq2Seq and transformer architectures like BERT and GPT, enhanced contextual understanding and query handling. Tools like SQLNet and TypeSQL further refine query generation by incorporating schema information. Hybrid systems combine neural models with rule-based logic for increased reliability, while interactive systems improve accuracy through real-time query refinement by asking clarifying questions. Additionally, commercial solutions like Microsoft Power BI and Google BigQuery offer text-to-SQL functionality, although they may not generalize well across different databases.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VII. FUTURE WORK & CONCLUSION

The Text-to-SQL Conversion System has the potential to revolutionize database interactions by allowing users to query in natural language. Future enhancements in Natural Language Processingwill improve understanding of complex queries and idiomatic expressions, while features like multilingual supportand machine learningwill broaden accessibility and adaptability. The addition of real-time feedback, contextual awareness, and visual query-building tools will enhance user engagement and simplify the querying process. Security measures will protect sensitive data, and integration with business intelligence toolswill facilitate efficient data analysis. Overall, these advancements will make the system a valuable resource across various industries, democratizing data access and empowering users to make informed decisions.

REFERENCES

1. Stinivasan, 1, Lei, J, Tau, Y., & Smith, E. (2023) butraction Tuning of Large Pre-trained Language Models. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), 1098-1107. https://doi.org/10.18653/v1/2023.acl- main. 109
2. Usta, A., Ozturk, O., & Yigit, E. (2022) DBTagger: A Deep Learning Approach ta Translate Natural Language Queries to SQL. In Proceedings of the International Conference on Data Engineering (ICDE), 456-465. https://doi.org/10.1109/ICDE 2022.456
3. Thome, 1, Williams, A., & Vlachos, A. (2022), NeuralDB: A Schema-less Database System with Natural Language Processing. Proceedings of the Conference on Empirical Methods in Natural Language Processing https://doi.org/10.18653/v1/2022.emnlp-main.285 (EMNLP), 3242-3253
MeCann, B., Raffel, C., & Socher, R. (2020), decaNLP: A Natural Language Processing Decathlon. Transactions of the Association for Computational Linguistics, 8, 422-438 https://doi.org/10.1162/tacl_a_00352
5. Chen, S, Xic, P. & Xing, E. P. (2021). Multi-Task Learning in Natural Language Processing: An Overview. ACM Computing Surveys (CSUR), 54(8), 1-28 https://doi.org/10.1145/3427714
6. Wong, A., Xu, J., & Li, Z. (2021) 4 Natural Language to SQL Model Based on the 75 Architecture. In Proceedings of the 35th AAAI Conference on Artificial Intelligence, 2045- 2052. https://doi.org/10.1609/anai.v35i3.1620
7. Naz, N. S., Hussain, T., & Khan, M. (2021). Automatic Correction of Semantic Errors in English Texts Using Weighted Federated Machine Learning. IEEE Access, 9, 138413138426. https://doi.org/10.1109/ACCESS 2021.3085072
8. Ekpenyong, M., Udo, S., & Edet, F. (2020). An Agent-Based Framework for a Natural Language Interface. Expert Systems with Applications, 162, 113751 https://doi.org/10.1016/j.eswa 2020.113751
9. Cheng, Z., Wei, Q., & Chen, Y. (2022). HiTab: A Hierarchical Table Dataset for Question Answering and Natural Language Generation, In Proceedings of the Conference on Natural Language Learning (CoNLL), 78-87. https://doi.org/10.18653/v1/2022.conllmain.8

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com