

AI POWRED EDUCAD USING SLM AND NLP

P. Aparna Devi¹, K. Shiva Kumar²

Department Of Computer Science and Engineering, Bapatla Engineering College, Autonomous and Approved By

AICTE-Accredited by NBA Affiliated to Acharya Nagarjuna University, Guntur, Bapatla-522102¹

Department Of Master of Technology in Computer Science and Engineering Bapatla Engineering College, Autonomous

and Approved By AICTE-Accredited by NBA Affiliated to Acharya Nagarjuna University, Guntur, Bapatla-522102²

Abstract: Natural Language Processing (NLP) systems enable machines to understand, interpret, and generate human-like language, bridging the gap between human communication and computer understanding. Natural Language Interface to Databases (NLIDB) and Natural Language Interface to Visualization (NLIV) systems are designed to enable non-technical users to retrieve and visualize data through natural language queries. However, these systems often face challenges in handling complex correlation and analytical questions, limiting their effectiveness for comprehensive data analysis. Additionally, current Business Intelligence (BI) tools also struggle with understanding the context and semantics of complex questions, further hindering their usability for strategic decision-making. Also, when building these models for generating the queries from natural language, the system handles only the semantic parsing issues as each column header is being changed manually to their normal names by all existing models which is time-consuming, tedious, and subjective. Recent studies reflect the need for attention to context, semantics, and especially ambiguities in dealing with natural language questions. To address this problem, the proposed architecture focuses on understanding the context, correlation-based semantic analysis, and removal of ambiguities using a novel approach. An Enhanced Longest Common Subsequence (ELCS) is suggested where existing LCS is modified with a memorization component for mapping the natural language question tokens with ambiguous table column headers. This can speed up the overall process as human intervention is not required to manually change the column headers. The same is evidenced by carrying out thorough experimentation and comparative study in terms of precision, recall, and F1 score. By synthesizing the latest advancements and addressing challenges, this paper has proved how NLP can significantly enhance the accuracy and efficiency of information retrieval and visualization, broadening the inclusivity and usability of NLIDB, NLIV, and BI systems.

Keywords: Correlation analysis, Enhanced Longest Common Subsequence (ELCS), Feature extraction, Semantic analysis.

INTRODUCTION

In today's data-driven landscape, efficient information retrieval and robust analytical capabilities are essential for informed decision-making across diverse domains. Data plays a pivotal role in driving the growth of businesses by enabling structured planning, informed decision-making, and effective execution. Decisions rooted in data have a higher probability of success, laying a solid foundation for business expansion and development. Utilizing tools such as business intelligence and data analysis empowers organizations to extract meaningful insights from their data reservoirs. Moreover, the use of natural language as input for generating insights is becoming increasingly significant. Natural language processing (NLP) techniques allow for the analysis of text-based data to extract valuable insights. By enabling users to pose queries in natural language, information retrieval systems can bridge the gap between users and data, facilitating more intuitive and efficient access to information. This integration of NLP capabilities enhances the usability and accessibility of analytical tools, empowering a broader range of stakeholders to derive insights and make data-driven decisions. In recent contributions, Natural Language Interfaces to Database (NLIDB) and Natural Language Interfaces to Visualization (NLIV) started using large language models (LLM) and semantic parsing approaches to understand the context of natural language effectively (Li et al., 2020b, Luo et al., 2021, Khan et al., 2024). However, as users are dealing with a variety of datasets with different contexts it becomes tedious to handle these NLP queries with minimal intervention in the database design. Existing systems might require manual interventions to identify the correct context, and correct abbreviation to avoid

ambiguity, this increases the burden on the DBA as well as the query processor (Yu et al., 2018b). Based on these observations we have identified the problems to be addressed in this research paper through our novel contributions. The

first major contribution deals with the correct context identification using correlation analysis. This correlation analysis processing can be used at different levels of hierarchies in businesses or applications where the impact of one feature/variable over the other plays a significant role in their decision-makings ([Anon](#), 2021). Existing Natural Language Interfaces to Visualization (NLIV) and Business Intelligence (BI) tools are not dealing with this problem efficiently. For instance, in education research, correlation analysis might explore the connection between study habits and academic performance, while businesses might investigate correlations between advertising spending and sales. To tackle this issue, we propose to combine correlation analytics with NL processing, aiming to reach the correct context with key features with the proper visualization which further generates the description which directly helps in effective query processing and decision making. The second contribution of this paper aims at handling ambiguities and abbreviations in tabular data. Real-time databases frequently employ abbreviations for column headers, leading to discrepancies between NL queries and table headers. This mismatch complicates the process of mapping NL queries to appropriate table headers, often necessitating manual intervention to align column headers with their full names. To address this challenge, the paper proposes a novel approach utilizing Enhanced Longest Common Subsequence (ELCS) to resolve ambiguities and abbreviations in column headers. By leveraging ELCS, the proposed method aims to facilitate seamless matching between NL tokens and table column headers, streamlining the process of mapping natural language queries to appropriate table headers. With these two contributions dealing with the existing issues, the paper aims to enhance the efficiency and effectiveness of natural language interfaces, enabling users to derive meaningful insights from NL queries and navigate ambiguities in tabular data more effectively. Ultimately, these advancements are expected to prove their effectiveness in decision-making, leading to optimized operational strategies and sustainable growth trajectories across various domains.

LITERATURE SURVEY RELATED WORK

In the current data centric environment, seamlessly interacting with databases and obtaining actionable insights through natural language has become critically important. Natural Language Interfaces to Databases (NLIDB) and Natural Language Interfaces to Visualization (NLIV) have been developed to facilitate this interaction, aiming to democratize data analysis by allowing users to query databases using everyday language. Despite significant advancements, these systems often struggle with complex queries and nuanced data interpretation. As businesses increasingly rely on data-driven decision-making, it is crucial to evaluate the effectiveness of these interfaces and identify areas for enhancement. This discussion delves into the capabilities and limitations of current NLIDB and NLIV systems, highlighting areas for future improvement and proposing solutions to enhance their effectiveness. Various approaches are available for creating a Natural Language Interface to database (NLIDB) ([Zhong et al., 2017](#), [Vaswani et al., 2017](#), [Xu et al., 2017](#), [Huilin et al., 2019](#), [Pennington et al., 2014](#), [Kim, 2014](#), [Yu et al., 2018a](#), [Yu et al., 2018c](#), [Li et al., 2020b](#), [Devlin et al., 2018](#), [Akhkhouk et al., 2021](#), [Li et al., 2020a](#), [Kummerfeld et al., 2018](#), [Dong and Lapata, 2016](#), [Zaremba et al., 2015](#), [Pal et al., 2019](#), [Kumar and Dua, 2014](#), [Bais et al., 2016](#), [Majhadi and Machkour, 2021](#), [Yu et al., 2018b](#), [Yu, 2021](#), [Brunner and Stocking, 2017](#)). Different techniques have been used to perform semantic analysis on natural language questions because of the ambiguous nature of natural language questions. Recently, large language models (LLM) like BERT, GPT, and T5 have been employed in NLIDB systems to improve the understanding of natural language queries ([Raffel et al., 2020](#), [Khan et al., 2024](#)). These models excel at capturing the context and nuances of language, leading to more accurate interpretations of user queries. However, these interfaces often fail to handle abbreviations and do not support analytical questions used to generate insights from data ([Sawant and Sonawane, 2023](#), [Miller, 1995](#), [Pedersen et al., 2005](#)). Natural Language Interface to Visualization (NLIV) systems indeed provide a user-friendly way for individuals to interact with databases using natural language questions ([Narechania et al., 2021](#), [Hu et al., 2019](#), [Luo et al., 2021](#), [Liu et al., 2021](#), [Luo et al., 2022](#)). However, they often fall short when handling complex analytical questions and providing detailed descriptions of visualizations. ncNet fails to predict the correct columns mainly because the NL query mentions desired columns ambiguously or implicitly. In some cases, it fails due to the “unperfect” table header (e.g., “StuId” vs. “Student ID”) ([Luo et al., 2022](#)). NL4DV does not work for vague predicates (e.g., ‘large’, ‘expensive’, ‘near’) ([Narechania et al., 2021](#)). Though Advisor can generate some visualization for the ambiguous question containing “trend” and “relations” but the model may fail for more complex questions ([Liu et al., 2021](#)). Additionally, these systems may struggle to generate descriptions of visualizations that effectively communicate the insights derived from such questions. Moreover, the lack of detailed descriptions accompanying visualizations can hinder non-technical users’ understanding and ability to communicate insights to higher management. Clear and comprehensive descriptions are essential for ensuring that business insights are effectively communicated and understood by all stakeholders. Business intelligence tools have indeed begun integrating natural language processing capabilities, but their ability to handle complex analytical questions still has limitations. While they excel at processing large datasets and providing insights, their capacity for handling intricate analytical queries remains a work in progress. Power BI and Tableau are the most well-known and widely used business intelligence tools ([Khatuwal and Puri, 2022](#)). They both offer powerful capabilities for data visualization,

analysis, and reporting, making them popular choices for businesses seeking to derive insights from their data. However, they still struggle with intricate questions requiring deeper semantic understanding.

Tableau's Ask Data feature lets users ask questions about their data in natural language and generates visualizations based on their queries, making it accessible for those unfamiliar with complex data analysis. Power BI's Q&A functionality also allows users to ask questions in natural language and receive visualizations in response, leveraging Azure Cognitive Service to understand and interpret queries. Both tools use natural language processing and AI-driven algorithms but struggle with complex queries like correlation analytics (Zhang et al., 2023).

PROPOSED SYSTEM

The proposed architecture aims to better understand and address such questions, bridging the gap between users and complex data analysis tools, and enabling a wider audience to derive insights without needing specialized skills in data visualization or analysis.

EXISTING SYSTEM

The expected outcome for the proposed algorithm is the matching table header. So the algorithm can successfully extract the matching table headers even if the abbreviations are being used in the actual datasets. The Enhanced LCS algorithm used finds the appropriate columns in less amount of time as memorization is being used along with the basic LCS. The time complexity can be reduced if there are overlapping subproblems, as the algorithm avoids redundant calculations by reusing previously computed results. As a result, the actual running time may be significantly less than $O(m * n)$ for certain inputs, especially if there are common subsequences within the input sequences. Additionally, for space optimization, only the previous and current rows of the matrix can be stored instead of storing all intermediate results in a dictionary. This approach reduces the space complexity from $O(m * n)$ to $O(\min(m, n))$, where m and n are the lengths of the input sequences. This optimization does not sacrifice the efficiency of the algorithm.

CONCLUSION AND FUTURE SCOPE

Natural language-based effective information retrieval has made significant strides, but there are ongoing challenges, such as contextual understanding and generating proper insights concerns. The proposed work is the first attempt to generate correlation analytics by understanding the ambiguous natural language questions for generating business insights. Hence it can be used to generate the proper visualization in the form of scatterplot, heatmap, etc. along with the description of visualization. The suggested methodologies can be added as an explicit phase to be executed into the existing analytical tools for converting ambiguous natural language tokens to correct table headers addressing correlation analytics without changing the column headers manually. The proposed framework with LCS ratio threshold given its best in terms of precision (99%), recall (97%), and F1 score (98%) for the expansion of abbreviations to correct table headers supporting the NL users in a better way. This can accelerate academic as well as industrial progress and help organizations make better decisions in less amount of time. Comparatively, suggested techniques outshine contextual understanding making them pivotal in shaping the future of information retrieval. Furthermore, a few more types of visualizations can be added depending on the use case.

REFERENCES

- [1]. Bhowmick et al., 2019 Bhowmick, R., Sadek Bhuiyan, M.I., Sabir Hossain, M., Hossen, M.K., Sadee Tanim, A., 2019. An Approach for Improving Complexity of Longest Common Subsequence Problems using Queue and Divide-and-Conquer Method, 1st International Conference on Advances in Science. In: Engineering and Robotics Technology. ICASERT, Dhaka, Bangladesh, pp. 1–5. <http://dx.doi.org/10.1109/ICASERT.2019.8934638>. Google Scholar
- [2]. Devlin et al., 2018 Devlin J., Chang M.-W., Lee K., Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding (2018), [10.48550/arXiv.1810.04805](https://arxiv.org/abs/1810.04805) View at publisherGoogle Scholar
- [3]. Dong and Lapata, 2016 Dong Li, Lapata Mirella Language to logical form with neural attention Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Long Papers) (2016), pp. 33-43, [10.48550/arXiv.1601.01280](https://arxiv.org/abs/1601.01280) View at publisherView in ScopusGoogle Scholar
- [4]. Ghayyur et al., 2018 Ghayyur O., Aleem M., Islam M.A. Time efficient novel parallel technique for Needleman Wunsch and Smith Waterman algorithms. International Conference on Computing, Mathematics and Engineering Technologies, iCoMET (2018), pp. 1-6, [10.1109/ICOMET.2018.8346356](https://doi.org/10.1109/ICOMET.2018.8346356) View at publisherView in ScopusGoogle Scholar