From Natural Language to SQL Query: A Medical Knowledge Graph-Linked Text-to-SQL Semantic Disambiguation and Enhancement Framework

Chongyun Zhou[1, 2], Yichen Bai[1, 2], Haiyan Li[1, 2, *]

[1]Institute of Information on Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing, China
[2]Library, China Academy of Chinese Medical Sciences, Beijing, China

Email address:

cmizcy@163.com (Chongyun Zhou), baiyichen2011@foxmail.com (Yichen Bai), lihy@mail.cintcm.ac.cn (Haiyan Li)
[*]Corresponding author

Abstract

Background: In the medical domain, efficiently retrieving structured data from large-scale databases using natural language queries remains a significant challenge. While medical knowledge graphs offer rich, structured representations of biomedical concepts and their relationships, a substantial semantic gap persists between these graphs and the free-form expressions used in natural language queries. Traditional Text-to-SQL systems often struggle with the complexity and ambiguity of medical terminology, leading to inaccuracies in generated SQL queries and hindering effective data retrieval and knowledge discovery. To bridge this gap, there is a growing need for approaches that integrate domain-specific knowledge with advanced semantic parsing techniques, enabling more precise and reliable translation of natural language into structured queries. *Objective:* To address the semantic gap between medical knowledge graphs and natural language queries, this paper proposes a Text-to-SQL semantic disambiguation and enhancement framework based on medical knowledge graph linking. The framework aims to improve the accuracy and reliability of converting natural language into structured query language, thereby supporting efficient medical data retrieval and knowledge discovery. *Methods:* First, entity linking and relation mapping techniques are employed to align medical entities in natural language with concepts in the knowledge graph. Next, a context-aware disambiguation model is introduced to resolve issues of polysemy and term ambiguity. Finally, a graph embedding enhancement module is incorporated to leverage path and structural information from the knowledge graph, thereby completing query semantics. Based on the structured semantic information derived from the previous steps, the final SQL queries are generated under the guidance of template-based prompts. *Results:* In experiments comprising 30 PubMed literature retrieval tasks, the proposed framework demonstrated complete consistency with human experts in terms of semantic coverage — successfully identifying and mapping all key entities and relationships from natural language queries to the knowledge graph. This represents a 12.32% improvement over the baseline model (DeepSeek R1) on this core metric. Regarding time efficiency, the framework completed all retrieval tasks in approximately 5 minutes, significantly outperforming manual retrieval by domain experts, which averaged 62.55 minutes. *Conclusion:* By deeply integrating medical knowledge graphs with semantic parsing techniques, this framework significantly enhances the semantic understanding capability and query generation quality of Text-to-SQL systems. It provides a reliable solution for medical knowledge-driven data retrieval, with potential applications in supporting clinical decision-making and scientific research analysis.

Keywords

Medical Knowledge Graph, Large Language Model, Semantic Disambiguation, Entity Linking, Query Enhancement, Clinical Data Retrieval, Biomedical Natural Language Processing, Semantic Parsing