International Journal of Data Warehousing and Mining
Volume 17 • Issue 2 • April-June 2021

Frameworks for Querying Databases Using Natural Language:
A Literature Review – NLP-to-DB Querying Frameworks

Syed Ahmad Chan Bukhari, Division of Computer Science, Mathematics and Science, Collins College of Professional Studies, St. John's University, USA

Hafsa Shareef Dar, University of Gujrat, Pakistan

M. Ikramullah Lali, University of Education Lahore, Pakistan

Fazel Keshtkar, Division of Computer Science, Mathematics and Science, Collins College of Professional Studies, St. John's University, USA

Khalid Mahmood Malik, Computer Science and Engineering Department, Oakland University, USA


https://orcid.org/0000-0002-7927-3436

Seifedine Kadry, Faculty of Applied Computing and Technology, Noroff University College, Kristiansand, Norway

ABSTRACT

A natural language interface is useful for a wide range of users to retrieve their desired information from databases without requiring prior knowledge of database query language such as SQL. The advent of user-friendly technologies, such as speech-enabled interfaces, have revived the use of natural language technology for querying databases; however, the most relevant and last work presenting state of the art was published back in 2013 and does not encompass several advancements. In this paper, the authors have reviewed 47 frameworks that have been developed during the last decade and categorized the SQL and NoSQL-based frameworks. Furthermore, the analysis of these frameworks is presented on the basis of criteria such as supporting language, scheme of heuristic rules, interoperability support, scope of the dataset, and overall performance score. The study concludes that the majority of frameworks focus on translating natural language queries to SQL and translates English language text to queries.

KEYWORDS

Cypher, Database, Natural Language Querying, NL2DB, NLP, NoSQL, SPARQL, SQL

INTRODUCTION

Several frameworks have been developed to translate natural language questions into a query language that can be executed over a database to retrieve the desired data. A key benefit of these translation frameworks is that they enable non-technical users to query database, without requiring any advanced knowledge of the query language syntax, as well as the design specification of a database (Reis, 1997; Christian, 2010). The history of natural language interface to database querying dates back to 1970s when the LUNAR and LADDER systems were developed for non-technical users to answer their questions about the moon rock samples and US naval ships, posed as natural language questions (Woods, 1972). The rapid evolution of computer hardware and software in the last five decades have

DOI: 10.4018/IJDWM.2021040102

led to such a revolution in such a way that the database systems which were developed in 1970s do not fulfill the current definition of a database system (Bercich 2003; Frank 2018). Ever since, several frameworks have been developed that translate natural language text to database query language. By studying the development timeline of such systems, we have identified interesting research trends in translating natural language to database queries domain. For instance, the CHAT-80 was the leading natural language to database query system which was developed in 1980 (Warren, & Pereira, 1982). Furthermore, early developed system had poor retrieval time, less support for the language portability, and had complex configuration processes. These factors contribute towards less adaptation of such systems for the commercial purposes.

Translating a natural language question into various database query languages such as SQL, Simple Protocol and RDF Query Language (SPARQL) is not a trivial task, as the current databases are diverse, gigantic in size, and follow sophisticated data storage mechanisms (Nadkarni, 2011). Storage engines often store data in a variety of ways such as in structured format (tabular), No SQL or graph (text) or in hybrid format. Therefore, underlying storage engines require different query languages to retrieve the stored data. This heterogeneity of data storage mechanisms increases the complexity of natural language to database query translation. With the advancement of machine learning techniques, various frameworks have been developed and are able to efficiently translate natural language questions (from simple to complex questions) into database specific queries (SQL, NoSQL) (Yossi Shani, 2016; Elías Andrawos, 2013).

The work presented in 2013 (Sripad and n.d. 2013) classified natural language querying framework for SQL only, according to the authors' knowledge, is the last published review paper on said area. Available review paper on this topic (Androutsopoulos, Ritchie, & Thanisch, 1995) has mainly covered natural language to SQL database and highlighted the usage of developed systems so far. In this survey paper, we have reviewed Natural language to database querying frameworks developed for both the structured (SQL) and non-structured database query languages (NoSQL, GraphDB). Using Google Scholar, we have found thirty-five relevant frameworks published from 2008 to 2018. This review excludes papers which describe proposed approaches without corresponding evaluation i.e. precision and accuracy, on any benchmark. We have sub-divided the developed frameworks into two main categories (SQL and NoSQL) and provided a comprehensive review of each section (Figure 1). Moreover, for each category, a feature comparison among the developed frameworks documenting their salient features and highlighting their shortcomings has also been provided. The comparison has been conducted on different factors including language and approach supported, performance evaluation and others.

SQL and non-SQL categories can be further divided into rule based and syntax analysis, syntactic pattern, machine learning and knowledge based/external resources. Furthermore, these sub-categories have been reviewed for different approaches including semantic matching, pattern matching, supervised and unsupervised learning and statistical approach. Statistical approaches use large text corpora and perform analysis based on text characteristics without considering significant linguistic knowledge. Similarly, symbolic approach is widely used as a learning measures to different machine learning techniques. Connectionist approach proves to be an efficient model of learning tasks, therefore, the combination of connectionist with statistical or symbolic approach is an important area in natural language processing (Stefan, Ellen, & Gabriele, 1996). Next section covers materials and methods used in conduction of this study.

Material and Methods

The most crucial part of this study was availability of relevant material. The articles were searched using authentic scientific databases including SPRINGER Link, IEEE, ACM Digital Library, Google Scholar, Emerald, Science Direct and Elsevier. Furthermore, some other databases were also explored but due to accessibility restrictions, they were not included. Search strategy was also designed based