

Natural Language Query System Frameworks: A Survey

¹Preeti Tuli, ²J.P. Patra

^{1,2} CSE Dept., SSIPMT, Raipur, India, preetituli@gmail.com

Abstract— A Natural Language Interface to Query Database enables users to query and retrieve information from a database without using any artificial language such as Structured Query Language (SQL). This paper presents a survey of natural language query system framework and aims at finding the issues in existing system, to address them in the proposed system. In the survey it has been found that 70% of the work in natural language interface to database querying has been carried out for SQL and NoSQL share 15%, 10% and 5% of languages like SPAROL, CYPHER and GREMLIN respectively. The Survey done claims that almost every frameworks implemented to work for English language. Thus, the proposed system can be implemented in the language other than English.

Keywords—NLI2QDB, IR, SQL, NoSQL, Database.

I. INTRODUCTION

Natural language to database querying frameworks translates natural language questions to valid database query languages. The translation acts as an interface between non-technical naive users to database systems, as users do not require to understand the database schema and query language syntax.

In this paper various research works have been studied to know about various frameworks already implemented and to identify the scope in the thrust area where few efforts have been put in.

In the field of natural language interface to query database NLIQD. It is observed that most of the available natural language to database querying frameworks support the English language only. Few efforts have been reported where researchers have worked on Portuguese and French. Multi-language support or dedicated systems in international languages are desirable to make the overall data-driven process easy[1].

II. METHODS AND PROCEDURE

Abhilasha Kate designed a system for Training and Placement cell officers who work on student database but don't have any knowledge of SQL. The System proposes the architecture for processing the English Query fired by T&P Officer to get SQL query using input as the text or speech. Speech recognition mainly deals with the ability of a machine or program to identify words and phrases in spoken language and convert them to a machine-readable format. The software has the ability to accept natural speech. Using

speech-to- text conversion will give the solution to the problems arising in the analysis or generation of Natural language text, such as tokenization, syntactic, semantic analysis, and usage of dictionaries and grammars necessary for such analysis. This translation is done by using Semantic Grammar [2].

TABLE I.
NLI2QD METHODOLOGIES IMPLEMENTED

Methodologies	Sample Work used Different Methods		
S.No.	Author/Paper title/Year	Algorithm/Method	Description
1.	Mathias Soeken et al. Automating the Translation of Assertions Using Natural Language Processing Techniques. FDL Proceedings ECSI 2014[3]	High level and low level abstraction using abstraction level assertions.	Author presented an algorithm that automates the translation of natural language assertions into SystemVerilog Assertions using natural language processing techniques. Instead of manually translating each assertion separately.
2.	Ryuichiro Higashinaka et al. Towards an open-domain conversational system fully based on natural language processing. Proceedings of COLING 2014[4]	Rule-based system.	This paper proposes an architecture for an open-domain conversational system and evaluates an implemented system. The proposed architecture is fully composed of modules based on natural language processing techniques.
3.	Anupriya et al. Fuzzy Querying Based on Relational Database. IOSR-JCE Jan-2014[5]	Fuzzy Logic	This paper mainly discusses the realization of fuzzy query through fuzzy theory and SQL combined C#. Also, a real life application
4.	Lei Zou et al. Natural Language Question Answering over RDF — A Graph Data Driven Approach. SIGMOD 2014[6]	Graph Data Driven Approach	Author proposes a semantic query graph to model the query intention in the natural language question in a structural way, based on which, RDF Q/A is reduced to subgraph matching problem. More importantly, author resolve the ambiguity of natural language questions at the time when matches of query are found. The cost of disambiguation is saved if there is no matching found

In earlier systems, while querying data from relational databases goes through one of the two ways: the keyword-based approach and the structured query approach. Both ways have their advantages and disadvantages. The structured query approach, while expressive and powerful, is

not easy for naive users. The keyword-based approach is very friendly to use, but cannot express complex query intent accurately. In contrast, natural language has both advantages to a large extent: even naive users are able to express complex query intent in natural language [2].

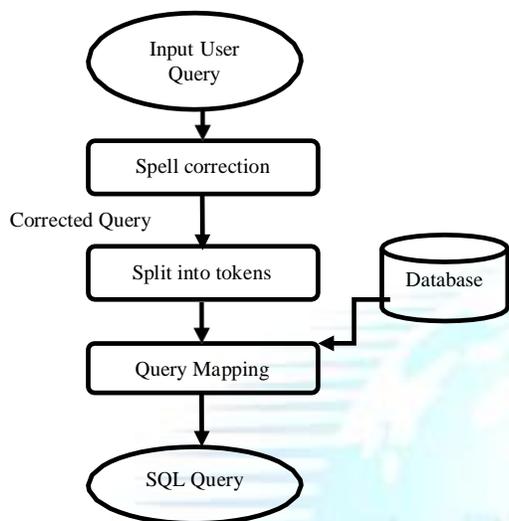


Fig.1 Earlier System

III. FRAMEWORK CLASSIFICATION

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, the symbolic approach is widely used as a learning measure to different machine learning techniques. Connectionist approach proves to be an efficient model of learning tasks, therefore, the combination of connectionist with the statistical or symbolic approach is an important area in natural language processing.

Hafsa et al. divided 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[1]

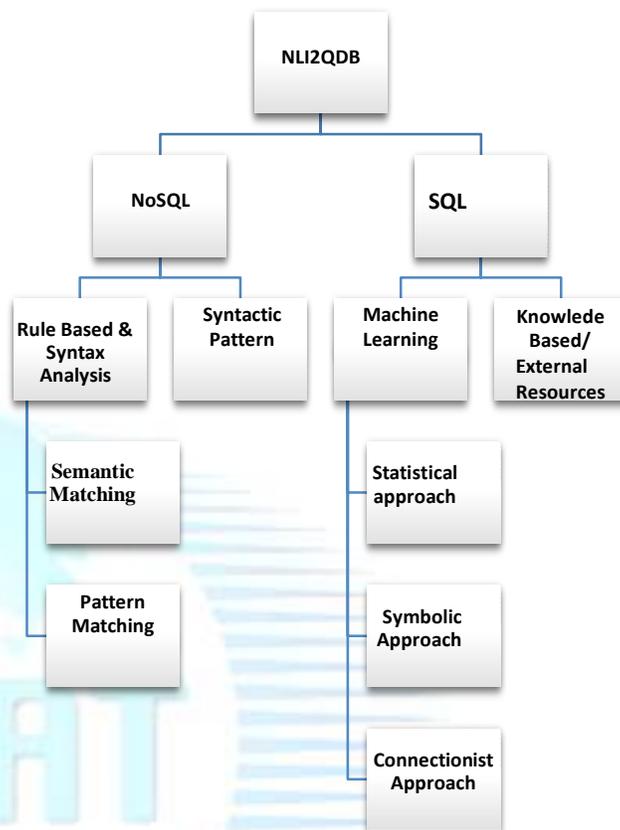


Fig.2 Classification of natural language to database querying frameworks.

There are many challenges in the conversion of natural language query to SQL query like ambiguity which means that one word can have more than one meaning. In this case, one word maps to more than one sense. Another challenge is the formation of complex SQL query and the next challenge is about Discourse knowledge in which immediately preceding sentence affects the interpretation of the next sentence for example if the user enters SELECT and INSERT query at the same time, then such a case is not understandable by the system.[1].

IV. CHALLENGES AND MOTIVATION

The conventional query in a relational database management system is not capable of satisfying the needs for dealing with queries that are in natural language. After going through different literature review, it is observed that earlier systems were not that much relevant to as per the user’s prospect. It condensed the user part of recalling the syntax but not to that amount. The system only provided a search interface for only partial database due to that user was obligatory to input queries according to that database. Through a literature survey some bottleneck is found in natural language query processing which is as follows:

1. The Difficulty of translating user-specified query structure to the actual schema structure in the database.

2. Spelling correction for mistakes made by the user while firing query maps the natural language query into database query language.
3. Semantic checking over user query.
4. The Most system works only for WH type question.
5. A limited data dictionary was used in which words need to be updated after a regular interval of time.
6. Natural language query had to be enclosed in double quotes
7. The Earlier system reduced the part of user but not to that extent that user required again he had to keep some things in mind that the name should in double-quotes, the query should consist of words that were present in the data dictionary.
8. Question words are also called WHQ type.

V. PROPOSED METHODOLOGY AND OBJECTIVE

The proposed intelligent information fetch system (NLI2DB) is based on natural language processing using fuzzy logic for complex query processing. In the proposed methodology a system is to be implemented as a software tool which can execute complex query on natural language, and we can fetch the information from the database effortlessly.

The main aspect of fuzzy logic is that it is able to deal with imprecise linguistic information which makes it attractive for automatic text summarization from the mathematical viewpoint; complex-valued fuzzy sets are natural. In classical binary two-valued logic, every statement is either true or false. On the computer, $-true$ is usually represented as 1, and $-false$ as 0. As a result, in the 2-valued logic, the set of possible truth values is a 2-element set $\{0, 1\}$.

The traditional two-valued logic is well-equipped to represent:

- situations when we are completely sure that a given statement is true.
- situations when we are completely sure that a given statement is false.

Initially in the proposed system, the input user query is taken as input in the database, the next step is to work on spell correction in sentence then the next steps are to split into the tokens after those next steps are semantic analyzer in the sentence. The next steps are query builder in the database and solve the query and finally find the result.

The corpus for natural language prepared, machine learning concept is proposed to apply for the preparation of corpus data.

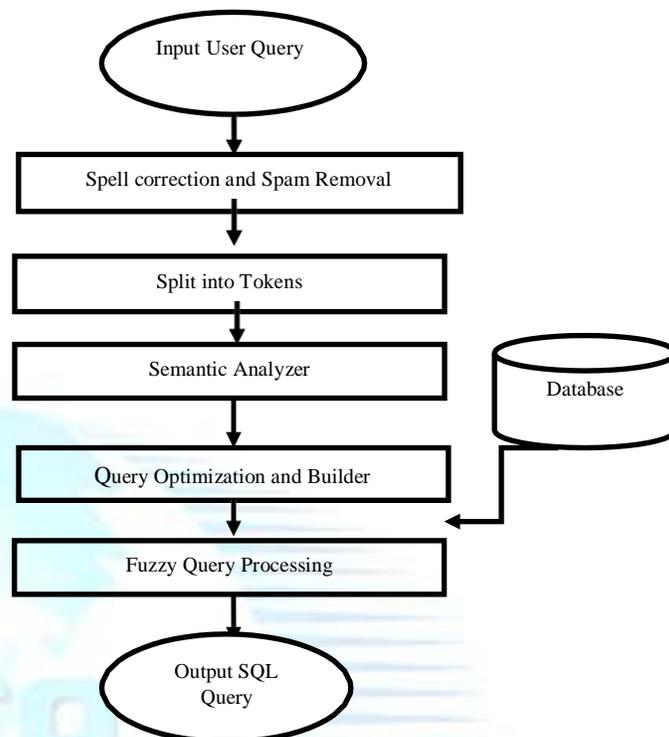


Fig.3 Proposed Methodology NLI2DB System

VI. CONCLUSION

Based on the study it is observed that the intricate and weak understand the semantics of documents and user queries has prepared feedback and alteration important distinctiveness of any IR systems. Hence, the natural language-based IR system will be much favorable. Natural language processing based IR systems is enormously capable to symbolize and manipulate the intricate query as complex and uncertain relationships presented among them.

1. The objectives of the proposed study are to aim at developing high-performance tool which will process natural language and convert it into SQL query.
2. The proposed method aims to remove spam from natural language queries.
3. Eventually will evaluate the accuracy of the proposed system.
4. The proposed system aims to handle joining queries.
5. The proposed system also aims to handle the slang.

REFERENCES

- [1] Hafsa Shareef Dar, M.Ikramullah Lali, Moin Ul Din, Khalid Mahmood Malik, Syed Ahmad Chan Bukhari, "Frameworks for Querying Databases Using Natural Language: A Literature Review", ArXiv, Corpus ID: 202541049, 2019.
- [2] G Abhilasha Kate, Satish Kamble, Aishwarya Bodkhe, Mrunal Joshi, Conversion of Natural Language Query to SQL Query, Proceedings of the 2nd (ICECA 2018) IEEE Conference Record # 42487.

- [3] Mathias Soeken, Christopher, Harris³ Nabila Abdessaied, Ian G. Harris, Rolf Drechsler, -Automating the Translation of Assertions Using Natural Language Processing Techniques¹, Department of Electrical and Computer Engineering, University of California, Irvine, United States.
- [4] Ryuichiro Higashinaka¹, Kenji Imamura¹, Toyomi Meguro², Chiaki Miyazaki¹ Nozomi Kobayashi¹, Hiroaki Sugiyama², Toru Hirano¹, Toshiro Makino¹, Yoshihiro Matsuo¹ -Towards an open-domain conversational system fully based on natural language processing¹, Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 928–939, Dublin, Ireland, August 23-29 2014.
- [5] Anupriya, Prof. Rahul Rishi, — Fuzzy Querying Based on Relational Databasel, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727 VolumIssue 1, Ver. I (Jan. 2014), PP 53-59.
- [6] Lei Zou , Ruizhe Huang , Haixun Wang , -Natural Language Question Answering over RDF —A Graph Data Driven Approach¹.
- [7] Tareq Abed Mohammed, Shavmaa Alavli, Saad Albawi, Adil Deniz Duru, Intelligent Database Interface Techniques using Semantic Coordination, 2018 IEEE, 1st International Scientific Conference of Engineering Sciences.
- [8] Anum Iftikhar, Erum Iftikhar, Muhammad Khalid Mehmood, -Domain Specific Query Generation from Natural Language Text¹, (INTECH) 2016 IEEE.
- [9] Mithun Balakrishna, Steven Werner, Marta Tatu, Tatiana Erekhinskaya, Dan Moldovan, K-Extractor: Automatic Knowledge Extraction for Hybrid Question Answering Mithun, 2016 IEEE Tenth International Conference on Semantic Computing.
- [10] Hanane Bais, Mustapha Machkour, Lahcen Koutti, Querying Database using a universal Natural Language Interface Based on Machine Learning, 2016 IEEE.

PRDG