

# AI-based Question Answering system for NoSQL standard query

Gourav Bathla<sup>1</sup>, Pardeep Singh<sup>2</sup> and Rahul K. Singh<sup>3</sup>

<sup>1,2,3</sup> *University of Petroleum and Energy Studies, Dehradun, India*

## Abstract

Question answering system is the significant application of Artificial Intelligence. If users ask for information from QA system in the form of natural language query, the syntax and semantics of query is understood by QA system and response is generated. Several existing solutions are proposed for AI-based QA system in E-commerce, Robotics, Autonomous vehicles and IoT-based solutions. However, to the best of our knowledge, very few research work have focused on QA system for NoSQL data stores. NoSQL (Not Only SQL) is data storage technique where Big heterogeneous data is saved by organizations. Researchers have proposed query languages for different types of NoSQL data storage. The limitations of existing solutions is that no standard query language is proposed and different query syntax are available for different data storage. If user asks in the form of natural language query, QA system should map this to respective NoSQL data store using Artificial Intelligence. In this paper, this issue is highlighted and key solutions are provided to resolve this issue. The key finding of our approach is that Question-Answering system interface is used where users can ask in natural language and query is mapped to specific NoSQL data store query using parse tree.

## Keywords

Artificial Intelligence, Question Answering, NoSQL, SQL, Natural Language Processing

## 1. Introduction

Data generated by social networking sites and real applications are unstructured and it is difficult for SQL-based system to process data [1]. Data storage techniques based on SQL are specifically designed for structured data i.e. which can be stored in rows and columns. In Big data, heterogeneous, flexible schema and horizontal scalability is required which is the reason NoSQL data storage techniques are proposed [2]. Companies are using NoSQL data storage to save large-scale data [3].

‘NoSQL’ term was introduced by Carlo Strozzi [4]. Column-Oriented, Document-based, Key-value and Graph-based are different categories of NoSQL data stores. Different NoSQL data storage techniques are used for specific purposes, for example, Key-Value data storage is used for get and put functions [5]. Object-relational mapping is a major issue in SQL [6].

Several researchers and practitioners have proposed query languages for NoSQL data stores. For example, CQL for Cassandra, UnQL for CouchBase, Cypher for Redis, MongoDB query language etc. The limitation is that syntax is different for these query languages. It is very difficult for data scientist to use multiple data stores due to several types of syntax and rules. There is need for standards of NoSQL query languages. In this paper, this standard is proposed using Artificial Intelligence based natural language query.

Several research works [7][8][9][10][11] have been carried out to map natural language to SQL queries but very few studies have focused on mapping natural language to NoSQL queries. Moreover, dependency parsing, Recurrent Neural Network, Seq2Seq, Attention-mechanism etc. are used for mapping natural language to SQL, but there is need for employing these techniques for mapping to No SQL data stores.

---

International Conference on Emerging Technologies: AI, IoT, and CPS for Science & Technology Applications, September 06–07, 2021, NITTR Chandigarh, India

EMAIL: gouravbathla@gmail.com (A. 1); pardeep.maan@gmail.com (A. 2); rahulsinghse25@gmail.com (A. 3)

ORCID: 0000-0003-4198-9647 (A. 1); 0000-0002-0368-4757 (A. 2); 0000-0002-4996-5300 (A. 3)

©2021 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



CEUR Workshop Proceedings (CEUR-WS.org)

The rest of the paper is structured as follows. In Section 2, NoSQL data storage categories are elaborated. Section 3 highlights NoSQL query languages and conversion of plain natural language to NoSQL data store query techniques. Finally, Section 4 concludes the paper with future directions.

## **2. NoSQL Data Storage**

In relational database, data can be stored in rows and columns format. Furthermore, entity integrity, referential integrity, joins, keys and normalizations are the essential features of SQL based databases. The difference in SQL database and NoSQL database is in supported data models and query languages [12].

In column-oriented data storage, data is stored column-based instead of row-based. Table's column reading is faster [13]. The main advantage is that dynamic schema is used in this data storage. There are a lot of sparse entries in row-oriented data stores as null value signifies the values that are unknown. In this data store, unknown values are don't included in flexible schema. HBase [14], Google BigTable [15] and Cassandra are examples of column-oriented database. In document-based data store, record is stored in document.

## **3. NoSQL query languages towards standard query language**

Query languages are useful for retrieving data of interest from data storage. Standards are developed for SQL queries and different categories are proposed. There is no standard benchmark for NoSQL queries.

### **3.1. CQL**

This language is for column-oriented data storage. This language is used for Cassandra. Column is the basic data structure in CQL [16]. It can manage structured and semi- structured data. Key spaces are used as database in CQL.

### **3.2. UnQL**

This language is specifically designed for Couchbase. Unstructured Query Language (UnQL) uses structural recursion to extract information from semi-structured data [17].

### **3.3. Cypher**

Cypher is graph-based language. It is originally implemented for Neo4j [18]. Cypher is well defined query language for processing graph-based databases. Create and select values from node can be easily implemented in Cypher.

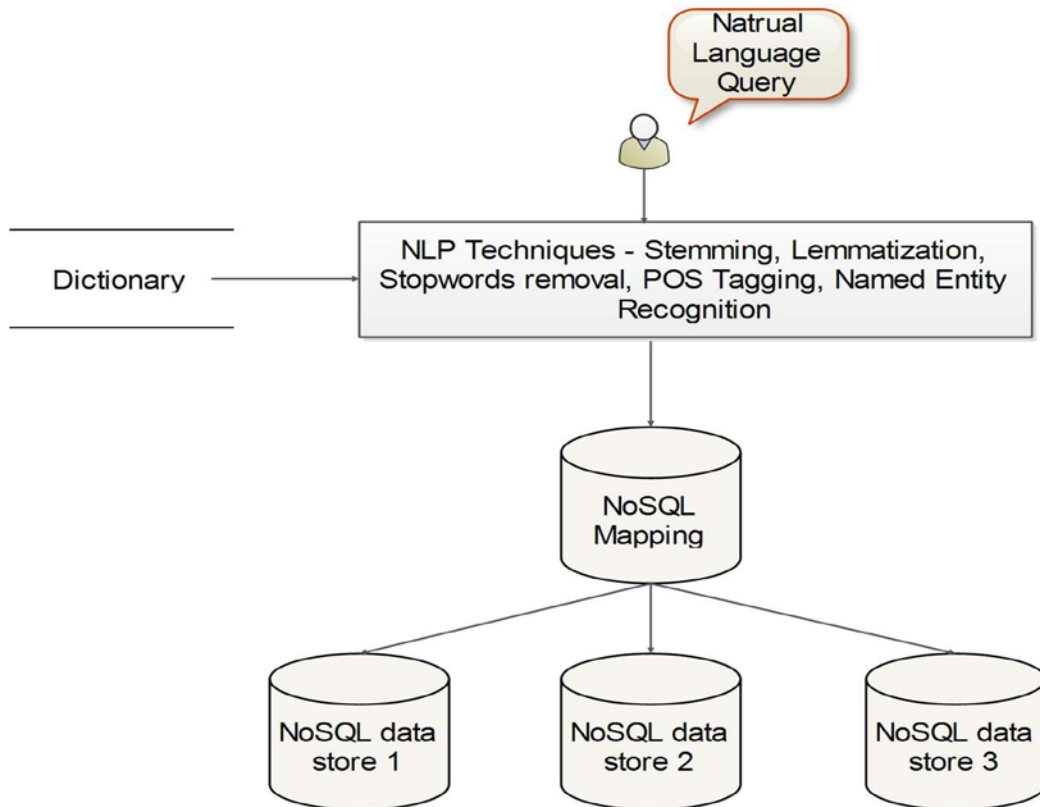
### **3.4. MongoDB Query Language**

MongoDB provides JSON based query language [19]. The difference between SQL query and this query language is that collections based queries are used for MongoDB. Several users want to migrate from SQL to NoSQL data stores as NoSQL data stores can handle unstructured data efficiently [24].

Users will just ask for data in plain natural language and further based on AI-based NLP techniques, query is mapped to specific database query. Natural languages are ambiguous so it is not easy for synthesizer [7]. Researchers have proposed approaches to convert natural language to specific data store query. For example, in [20], Natural language query is translated to document-based query. In [21] question-answering system is developed to query NoSQL database. SQL queries

are automatically synthesized using natural language in [7]. Natural Language query is mapped to MongoDB query in [22]. Natural language to NoSQL query is generated using query response model in [23].

In Figure 1, it is clearly depicted that users need to fetch the record using natural language query. This query is preprocessed to remove stopwords and use stemming and lemmatization. Then NoSQL mapper converts the natural language query into specific NoSQL data store query. It is forwarded to particular NoSQL data store to retrieve values. If data scientist needs to fetch record of details of product ID 101, Query in natural language will be ‘What are the features of product id 101’.



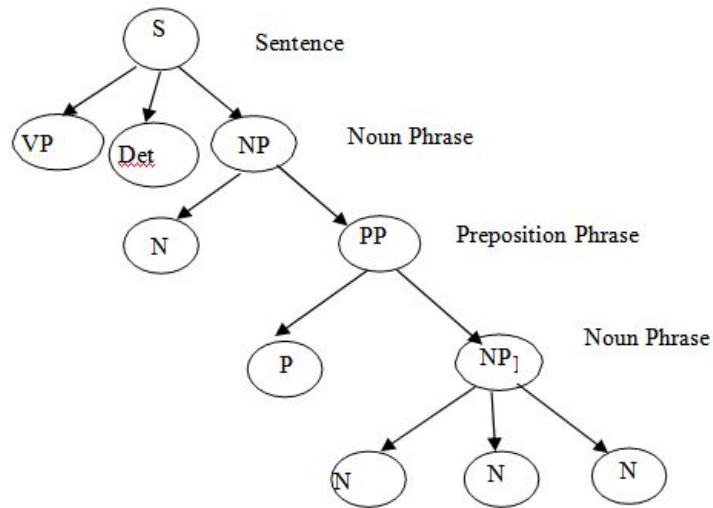
**Figure 1:** Natural language query to specific NoSQL query

‘What’ is used in natural languages so it is concluded that user wants to select data, so ‘select’ query will be used. ‘feature’ and ‘id’ are used as nouns, so user wants to ask about features. ‘of’ is used as preposition and after this ‘product’ is used as noun, so it is clear that user wants to ask about product id 101. POS tagging will be - ‘What’ and ‘are’ -Verb Phrase, ‘feature’ – Noun, ‘of’ – preposition.

It is depicted in Figure 2 that parse tree is created for query asked in natural language. This parse tree is used to map this query to NoSQL data stores query. ‘S’ as complete sentence is parsed and it derives VP and NP, where VP is combination of ‘what’ and ‘are’ from which it is concluded that ‘select’ has to be used for database query.

It will be further mapped to: (a) SQL: ‘select \* from product where id= ‘101’ (b) MongoDB: ‘db.collections.find (‘id=101’) (c) CQL: ‘select \* from product where id=101’ (d)Cypher: Match (E: Employee {id: 101}) return E

It clearly validates that natural language query can be mapped to various NoSQL query languages provided stemming, lemmatization, POS Tagging, Named Entity Recognition and semantic analysis is deployed on natural language query.



**Figure 2:** Parse tree for Natural Language query

#### 4. Conclusion and Future Work

In this paper, Question Answering system is proposed for mapping natural language query to respective NoSQL data stores. User will ask query to QA system in English like language and using Artificial Intelligence application, query is mapped to NoSQL database using parse tree. This approach will be useful for NoSQL-based organizations as there will be no need to propose QA system for different NoSQL data stores. Our proposed approach will provide comprehensive structure to use any NoSQL database. In this paper, NLP which is significant application of AI is used to propose QA system. In future, different evaluation metrics and deep learning will be used to analyze accuracy of mapping from natural language to NoSQL query language.

#### 5. References

- [1] S.Ghotiya, J.Mandal,S.Kandasamy, Migration from relational to NoSQL database. In IOP Conference Series: Materials Science and Engineering. IOP Publishing,2017, pp. 1-7.
- [2] G.A. Schreiner,D.Duarte, R. dos Santos Mello, Bringing SQL databases to key-based NoSQL databases: a canonical approach. Computing 102 (2020), 221-246.
- [3] F.Michel, C.Faron-Zucker, J. Montagnat, Bridging the semantic web and NoSQL worlds: generic SPARQL query translation and application to MongoDB. In Transactions on Large-Scale Data- and Knowledge-Centered Systems XL (2019) 125-165.
- [4] C.Strozzi, Nosql: A non-sql relational database management system, 2013.
- [5] F.Gessert, W. Wingerath, S.Friedrich, N.Ritter, NoSQL database systems: a survey and decision guidance. Computer Science-Research and Development 32, (2017), 353-365.
- [6] K.Kaur, R. Rani, Modeling and querying data in NoSQL databases. In 2013 IEEE International Conference on Big Data, 2013, pp. 1-7.
- [7] N. Yaghmazadeh, Y. Wang, I.Dillig, T.Dillig, SQLizer: query synthesis from natural language. Proceedings of the ACM on Programming Languages, 2017, pp.1-26.
- [8] H. Kim, B.H.So, W.S. Han, H.Lee, Natural language to SQL: Where are we today?. Proceedings of the VLDB Endowment,2020, pp.1737-1750.
- [9] S. Yavuz,I.Gur, Y. Su, X. Yan, Dialsql: Dialogue based structured query generation. In ACL, 2018, pp. 1339–1349.
- [10] S.Iyer, I.Konstas, A.Cheung, J.Krishnamurthy, L. Zettlemoyer, Learning a neural semantic parser from user feedback. In ACL, 2017, pp. 963–973.

- [11] V. Zhong, C.Xiong, R.Socher,Seq2sql: Generating structured queries from natural language using reinforcement learning. arXiv preprint arXiv:1709.00103,2017.
- [12] S. Bjeladinovic,Z.Marjanovic, S. Babarogic, A proposal of architecture for integration and uniform use of hybrid SQL/NoSQL database components. Journal of Systems and Software (2020), 110633.
- [13] D.J.Abadi,P.A.Boncz, S.Harizopoulos, Column-oriented database systems. Proceedings of the VLDB Endowment, 2009, pp.1664-1665.
- [14] HBase, <http://hbase.apache.org>
- [15] F. Chang, J. Dean, S. Ghemawat, WC. Hsieh, DA. Wallach, M. Burrows, T. Chandra, A. Fikes, RE. Gruber, Bigtable: a distributed storage system for structured data, ACM transaction on Computer Systems 26,(2008).
- [16] J.Carpenter, E.Hewitt, Cassandra: the definitive guide: distributed data at web scale. O'Reilly Media,2020.
- [17] P. Buneman,M. Fernandez, D. Suciu,,UnQL: a query language and algebra for semistructured data based on structural recursion. The VLDB Journal (2000), 76-110.
- [18] N. Francis,A. Green, P. Guagliardo, L.Libkin, T. Lindaaker, V. Marsault, S. Plantikow,M. Rydberg, P.Selmer, A. Taylor, Cypher: An evolving query language for property graphs. In Proceedings of the International Conference on Management of Data, 2018, pp. 1433-1445.
- [19] F. Michel,C.Faron-Zucker, J.Montagnat, , September. A mapping- based method to query MongoDB documents with SPARQL. In International Conference on Database and Expert Systems Applications, 2016, pp. 52-67.
- [20] T. Pradeep, P.C.Rafeeqe,R. Murali, Natural Language To NoSQL Query Conversion using Deep Learning.
- [21] S. Blank,F.Wilhelm, H.P.Zorn, A.Rettinger, Querying NoSQL with Deep Learning to Answer Natural Language Questions [in press]. In Conference on Innovative Applications of Artificial Intelligence,2019.
- [22] M.D. Gadekar,B.M. Jadhav,A.S.Shaikh, R.B.Kokare,Natural Language (English) To MongoDB Interface. International Journal of Advanced Research in Computer Engineering & Technology (2015).
- [23] S.Mondal, P. Mukherjee,B.Chakraborty, R. Bashar, Natural Language Query to NoSQL Generation Using Query-Response Model. In 2019 International Conference on Machine Learning and Data Engineering, 2019, pp. 85-90.
- [24] H. Matallah,, G.Belalem, , K.Bouamrane, Comparative Study Between the MySQL Relational Database and the MongoDB NoSQL Database. International Journal of Software Science and Computational Intelligence (2021) 38-63.