A Review on Information Retrieval – Natural Language Processing Approach

H. C. Vijayalakshmi
Department of Computer Science
JSS Science and Technology University (Formerly SJCE)

Manasamithra P.
Department of Computer Science
JSS Science and Technology University (Formerly SJCE)

ABSTRACT

Information retrieval is very important area in any of the IT applications. Systematic data storage is essential in information retrieval. In conventional method, data is stored in a structured format and retrieved using SQL queries which requires technical knowledge. Most common data storing and retrieving mechanism in IT is the Relational Database Management Systems. Now there is a necessity to retrieve data from a very large database using natural language from any social media. A brief survey of various methods used to store and retrieve data using Natural Language Processing (NLP) is carried out in this paper.

Keywords

Big data, Database Management System, Information retrieval, natural language queries, SQL, Unstructured data

1. INTRODUCTION

SQL is the standard query language for relational databases used for retrieving information from relational database where data exists in a structured format. People without technical background find it difficult to retrieve information from the database using computer languages. It is also challenging for the SMEs (Subject Matter Expertise), because it requires the users to have knowledge about schema of the database. It is very useful to have a user interface for retrieving the information. Nowadays, very huge amount of information is available in an unstructured format (face book, Twitter, e-commerce sites etc). Many industries extract useful information from this already available big data and leverage the same for some meaningful application.

Natural language processing is one of the emerging technologies which make the user computer interaction easier. User can query the database using natural language and system should be able to answer intelligently and accurately. This is the core concept under Natural language processing. The study of approaches underlying this is very important to understand the emerging technologies and to improve the shortcomings of the existing systems. Various techniques are proposed by various researchers and experimented on different data sets and computed the accuracy. This paper does a systematic review on various methods and approaches to develop such interfaces for both structured and unstructured data.

This paper is organized as follows. Section 1 gives brief introduction about different methods proposed and used. Section 2 gives a detailed description of various existing system architectures, Section 3 briefs about Stanford Dependency Parser which is used for parsing the query posed by the end user in natural language. In section 4 a comparative analysis of various NLP techniques proposed by different scientists in the area of unstructured data is discussed. Section 5 gives a detailed analysis of our overall study followed by Conclusion and Future Work in Section 6.

2. DETAILED DESCRIPTION

In [1] Fei L et al. aimed at constructing an interactive natural Language interface for relational Databases. Dependency parser is used to understand the natural language query linguistically. Here Stanford parser is used to generate dependency parse tree. The parse tree node mapper identifies the nodes in the linguistic parse tree. Then it is mapped to SQL components and tokenizes them into different tokens. For each such node, the parse tree node mapper outputs the best mapping to the parse tree structure adjustor by default and reports all candidate mappings to the interactive communicator. Parse Tree Structure Adjustor step is to correctly understand the tree structure from the database perspective. Implicit nodes are inserted to include it under the syntactic coverage. Once the query tree is verified by the user, it is converted into SQL expression with appropriate PK-FK relations and joins.

The system is almost accurate if the user provides the correct verification during interaction. But the disadvantage is that user should have minimal knowledge of interpreting the query tree. The system overview is shown in Figure 1.

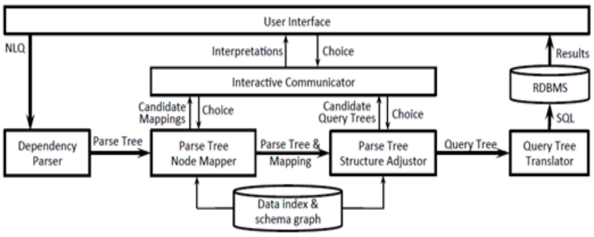

Figure 1: System architecture Courtesy Fei L et al. [1]

The system proposed by Garima Singh et al [2]. uses an algorithm based technique to generate the SQL query from the natural language. The system analyses Natural language query in series of steps and at each stage the data is further processed to finally form a query leading to its execution. The first step performed is Tokenization where the sentence is divided into smallest possible words called tokens. Next is the Escape word removal where the stop words and words which do not contribute anything in forming the SQL query are removed. This is followed by identifying the Part of Speech Tagger. i.e., the tokens are classified into nouns, pronouns, verbs and string/integer variables. Finally, system classifies the tokens into relations, attributes and Clauses System architecture in shown in Figure 2. Also, it removes the ambiguity if there any. Final query is constructed from inputs given by all the previous steps. The data fetched from the database according to the query formed. Accuracy is less due to limited set of pre-defined rules which are used in implementation.

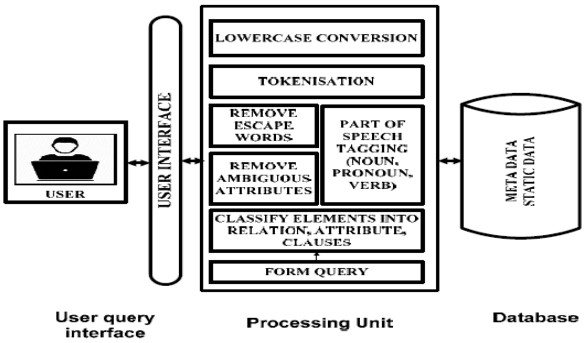

Figure 2: Three tier Architecture of the system. Courtesy Garima Singh et al [2]

Subhabrata Sengupta et al. [3] have described a system, which processes the natural language by using a method called “Levels of language” also known as synchronic model of language. The steps involved in this model are namely morphological analysis, Lexical analysis, Syntactic analysis, Semantic analysis and Speech to Text Recognition. Here Python for Android is used for Text Recognition.

The tool takes the input in the form of either speech or text. Convert the speech to text and then split the query into tokens. Find all the attributes and tables present in the query. This is followed by identifying the attributes which belongs to the database table. Form the simple query by identifying the conditions present. Further look for the additional tables if natural join is present.

This system is capable of handling simple queries along with some complex queries, but not all forms of SQL queries. Also, this system is tested on a small college database hence the result on large dataset is unknown. Success rate is less since it gives wrong results for syntactically incorrect sentences.

Prabhdeep Kaur et al. [4] have used the different phases of the language processing to create the SQL query. The input is the speech which is converted to text. The text is then parsed in different stages to produce the final result. This system works only for simple queries. “GROUP BY” and “HAVING” are not supported here.

Axita Shah et al. [5], overcame the disadvantage of [3] by providing correct result for syntactically incorrect sentences also. The first part of the system is the Natural Language Interface to Database(NLIDB). This approach has three components, namely Language Component, Intermediate Language Representation component and Database Component. Language component contains morphological, syntactic and semantic analysis. One enhanced technique used in the syntactic analysis is the stanford Parser to generate the syntax tree which increased the accuracy. Intermediate Language representation provides knowledge base to the Database component. Finally, DB component generates the SQL query using knowledge of IR component. Second method used is the Keyword based Interface to Database(KBIDB). It uses the DBXplorer to search in relational databases using two preprocessing steps called publish and search. It gave very good accuracy of 53% in case of syntactically invalid natural language queries.

Filbert Reinaldha et al. [6], proposed an advanced technology by which Natural language interface to the relational database was able to solve question type queries and unit conversion. Data conversion includes measurement units and some currency units. Question analyzer uses Stanford Dependency parser to obtain the dependency between the words which helps in translation process. Query generator analyzes dependencies and converts to SQL query. Metadata of the database is used to create the ontology. Ontology is used o identify the objects in the user input. The types of questions handled are wh-questions, yes-no questions and tag questions. The accuracy is very high except in the cases where stanford dependency produces the wrong result. This paper develops a NLIDB system that uses ontology as semantics processing and Stanford Dependency Parser as query analyzer in processing question type query.

Azilawati Azizan et al. have used an approach in [7] which deals with the application of information retrieval in search engines. The ontological approach is used in the document search. In this paper, a comparative analysis of using ontology and without using ontology in document search. Researchers have used Durian Ontology. The methods used for implementation is the combination of keyword based and ontology. The result obtained is fairly accurate.

Manavalan et al. [8], have proposed four different algorithms to process the user query based on the subject. Also, the dataset used is slightly different. The data lies on different geographical areas and in the virtual machines of different servers. Based on the user input, system will decide on algorithm to be selected for processing. Criteria for deciding the algorithm is based on the attributes present in the VMs. The algorithms are for processing the different conditions like natural join, group by or other type of conditions. Even though the results are encouraging, the system is little complicated to understand from the end user perspective.

In [9] Sanket S.Pawar et al., used the concept of web search and extended to Information retrieval in case of relational databases. It requires extending Index, ranking, clustering to database management system. This system is especially built on the keyword based technique. The name, two-way views is given, since two techniques are compared. One is IR style system and the other is, Discover Approach with candidate Network. It is a keyword based approach. The modules of the system are Query processing where structured data is formed, Tuple set Extraction in which the search based tuples are extracted from the DBs, IR Engine Candidate network generation where all mappings of the tuples are done from different keyword trees and Performance evaluation. It often gave the average result on selected dataset. But execution time expensive in case of keyword based search.

The new concepts in [10] suggested by Xuan Xuan et al., are that it used the Chinese language as the natural language to query the system. Database used for the testing is remote sensing data which contains some conceptual data for which general NLP techniques are not sufficient for the purpose of mining. Pattern matching, Dependency Extraction and Knowledge Extension are used in addition to the NLP. Again, it is a keyword based approach, so the result is same as other keyword based techniques mentioned above.

Xu Yiqiu et al. [11] proposed different levels of language components such as lexical analysis, syntax tree, and semantic grammar followed by an intermediate language representation. Natural language query sentences are translated into intermediate language query sentences, and then converted into SQL query sentences. They have also used Stanford parser for understanding dependency. Types of queries supported are Yes/No Query, when queries, Command Query and hybrid queries which is a combination of all the above. Semantic interpreter plays the main role in deciding the place holder of each word from the natural query. This novel initial experiment gave out a good result.

Mahesh P.Gaikwad et al., provide a slightly different approach of natural language processing technique in [12]. The query finding algorithm is different than others. The finding of the different components of the SQL query is carried out in certain steps. If the word does not belong to one category, then only the algorithm will move to find out the next category. This is elimination technique where initially find out all the table names. Then ignore those words for next level search of attributes.

In addition to the features built in previous systems, Pooja A.Dhomne et al in [13] has an option to input our natural query in two languages. Either English or Marathi is used. Query given is parsed to produce parse tree which is then interpreted semantically. Then it is represented in an intermediate language. Finally, data is retrieved using the produced SQL query. Again, this approach is common in almost all the implementations which give almost average result.

Avinash J. Agrawal et al. proposed a system in [14], which is purely ontology based. Here the natural language query is processed by chunking the similar group of words. Tag the Part Of speech and then identifying the named entities. Then the deep semantic analysis is done by forming a domain based ontology which contains all the objects, its properties and constraints. Map the ontology with the preprocessed query. It results in the correct position of the words in the sql query. For eg, if the question starts with ‘what’, the noun phrase immediately after the wh-word determines the expected entity. Result obtained was 90% accurate. It was good compared to other papers.

Rukshan Alexander et al. [15] designed a Natural Language Web Interface for Database which enables the users to interact with the system over the internet without any constraints on the domain. It also extends the study to Question & Answer system. (Q&A). The method used is, after preprocessing the query, SQL elements are identified and compared with already prepared SQL templates. Based on the matched template, SQL query is generated.

In [16], Gaganpreet Kaur et al. used the concept of regular expression in conversion of natural language query to SQL. Regular expression can be generated from a language by creating finite state automata. Multiple sequence alignment is an improved algorithm which is proposed to generate more suitable regular expressions for NLP. A particular RE constraint is followed in this approach. Another method to generate RE is, Automated Discovery of Valid Test Strings from the Web using Dynamic Regular Expressions Collation and Natural Language Processing, which gives realistic result. ‘StriDFA’ is another approach for multi string matching to find out more accurate Regular Expressions, where multi string matching can be performed. This paper mainly concentrated on the preprocessing and preprocessing result was accurate.

Akshay G. Satav K et al. [17] used a simple approach. Here input natural language query is broken into meaningful tokens. Then, these tokens are put in appropriate place in the SQL syntax using mapping rules.

3. STUDY ON STANFORD PARSER

Many Natural Language processing implementations used Stanford dependency parser as the syntactic and semantic analyzers. The Stanford typed dependencies representation is designed to provide a simple description of the grammatical relationships in a sentence that can easily be understood and effectively used by people without linguistic expertise who want to extract textual relations discussed by Valentin Ilyich Spitkovsky et al. [28]. In particular, rather than the phrase structure representations that have long dominated in the computational linguistic community, it represents all sentences relationships uniformly as typed dependency relations which called ‘typed dependency parser’ and proposed by Marie-Catherine de Marneffe in [29]. Typed dependency grammar gives the part of speech tagging to each word present in the sentence.

Example: Tree representation for the query - Return all the authors who have more papers than Bob VLDB after 2000

The resultant dependency parse tree is shown in Figure 3.

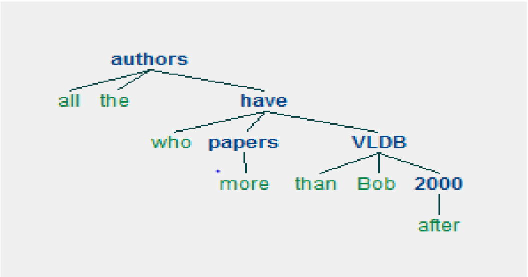

Figure 3: Stanford Dependency Parse tree.

4. COMPARATIVE ANALYSIS

All the papers above described the methods of retrieving the information from relational database. This section discusses various methods of retrieving information from unstructured data. A comparative analysis of various proposed methods by different scientists is tabulated which includes the paper name, Dataset used in the implementation, brief explanation of the method, Execution time, Accuracy and is there any scope for improvement sections.

Javubar Sathick et al. [18] focused on creating SQL query for extraction of knowledge from web. Initially capture the data from the web in raw format. Then it is converted into structured format and stored in database. R Studio is used in this implementation. This is followed by regular mining having four main components. First is the morphological analysis where the sentence is broken down into tokens. This is followed by stemming. Semantic Analysis finds the relationships between the words, by forming parse tree. Then a set of sql templates will be saved as mapping source. Once the parse tree is obtained, keywords which fits into ‘select’, ‘from’ and ‘where’ clause are identified. Based on the mapping, these words are kept in the corresponding template. This gave rise to a new concept where it is possible to retrieve data from unstructured dataset like twitter, facebook etc. This is an initial effort of mining unstructured data by converting it into structured. So, it takes lot of effort in conversion.

Rongrong Zhang et al. [19], presented the application of understanding the short domain question in natural language to query the data. A data dictionary is formed initially by extracting schema of all tables and short text data from the database using ODBC API. This is basically a question and answer system which extracts the keywords using an IK Analyzer segmentation using domain dictionary. This is followed by answer retrieval from the database using converted SQL. Even though this is a relational database concept, it is extended to Q&A system which can be used in retrieving social media data.

Johanna Monti et al. [20], puts an effort on Cross Language Information retrieval application where any language query is translated to a common language say English, and retrieve the information from web. The methodology is based on the Lexicon Grammar which encompasses both syntax and lexicon. Semantic information is stored in electronic dictionaries and Finite Automata. Development of architecture with a central multilingual formalization of the lexicon and extraction ontologies and SPARQL/SERQL adaptation systems increases the accuracy of the Q&A systems. This relatively is a new approach to process big data.

Mochamad Vicky Ghani Aziz et al. in [21] have used natural language processing to extract the traffic information from social media data. Twitter is used as the sample data. This is implemented in four stages namely, data collection, preprocessing, natural language processing and classification of traffic conditions. In the NLP stage, POS tagging and syntax analysis is done on the cleaned keywords. Finally classify the traffic in the form of a table from the previous steps result. This gave a different application of natural language processing, but the accuracy is too low.

In [21], the result was not accurate due to poor preprocessing technique. Belainine Billal et al. in [22] dedicated to a different methodology for preprocessing the big data such as twitter data so that the processed data can be useful for natural language processing. This is made up of pre-processing pipeline for tweets consisting of filtering part-of-speech, Named Entities Recognition (NER), hash tag segmentation and disambiguation. NER is the process of identifying real world entities such as person or place name. Hash tagged words are usually the words which explains the context which can be identified through use of regular expressions. Word disambiguation is done by using wordnet which provides a synset of all the synonyms. Then using various pattern matching techniques such as PCA, Naïve Bays method of classification, tweets can be classified successfully and can be used for further processing.

Abdeljalil el abdouli et al. in [23] proposed an approach for analyzing the data generated by Moroccan users in the social network Twitter, in order to discover the subjects that Moroccan society is interested in and then locate on Moroccan map the areas from where the tweets come from. Main concept in any machine learning project is the method of storing the data. A distributed file system called HDFS (Hadoop Distributed File System) is used here to store the tweets. Raw tweets are analyzed by using following 3 approaches. MapReduce of Framework Apache Hadoop, Python language for implementation and Natural Language Processing (NLP) techniques. The corpus then generated into numeric features and k-means algorithm is applied to cluster all words into general topics. Finally, tweets are plotted on Moroccan map by using the coordinates extracted from them, in order to have an idea about the geo location of these subjects.

Mohd Suhairi Md Suhaimin et al. [24], proposed a process in which first extract the features using NLP, then identify the sarcasm in the context of bilingual data. The process of feature extraction consists of two main steps: (i) extraction of the lexical, pragmatic and Malay prosodic features from the original bilingual corpus and (ii) translation of the bilingual corpus to English and extraction of English prosodic features, along with syntactic and idiosyncratic features. This process returns a set of features namely, Lexical, Pragmatic, Prosodic, Syntactic, Idiosyncratic features. Economic news from Facebook public pages is used as the dataset.

Li Xin et al. [25] used a approach where data is extracted using web crawling algorithm. Firstly, the natural language processing module in Thomson Data Analyzer is applied to extract the keywords from the tweets. Secondly, the social awareness information is analyzed by applying text mining and social network analysis, then, the social awareness of emerging technologies is mined by using time-slicing-based awareness information map. Finally, the Perovskite solar cells technology is as a case study to analyze the effectiveness and feasibility of the method.

In [26], Kaixian Yu et al. proposed Grammatical Relationship Graph for Triplets (GRGT). Here the relationship between the words of interest is extracted. Words are extracted using Natural language processing techniques and relationship is found out using graph theoretic algorithm. First of parse the given natural language sentence by using parsing methods which results in grammatical graph. The graph is further divided into sub-graph if needed. Use the shortest path distance algorithm to find out the relationship. This gave a good result on biomedical data than the best performing method in literature. But the scope is limited to only bio-entities.

Ranjeet Devarakonda et al. [27] has discussed about a system in which natural language processing is used to make the data of the Atmospheric Radiation Measurement Data Center (ADC). It can be more reachable to public. The architecture uses Apache Lucene/Solr, OpenML, and Kafka to generate an automated query/response system with inputs from Twitter, Cassandra DB, and log database. Data producer and consumer workflow using Kafka is the main technique used here.

In [28], Donia Gamal et al. focused on opinion mining using Machine learning and Natural language processing. Several algorithms such as Support vector machines, Naïve Bays, maximum entropy is utilized to extract the information and differentiate the user’s opinion. The opinion is categorized into positive, negative and neutral. From the study of the author, it is clear that SVM with POS gave the highest accuracy of 92%.

A comparative analysis of different NLP techniques used by different researchers is tabulated in Table 1.

Table 1. Comparative Analysis of information retrieval techniques

Sl No	Paper Name (Authors)	Dataset used	Approach	Result Accuracy	Scope for Improvement
1.	Fei L et al. [1]	Microsoft Academic Search (MAS)	Linguistic Parse Tree from the Stanford Parser	90%	Yes
2.	Garima Singh et al. [2]	Education Institution Database	Predefined structures are employed, and the system is trained using NLP rules. Keyword based approach.	82.1%	Yes
3.	Subhabrata Sengupta et al. [3]	College database	Processing the Natural language based on the Levels of Language known as Synchronic Model of language	Limited Accuracy	Yes
4.	Prabhdeep Kaur et al. [4]	Sample Employee Database	Different stages of natural language processing	Limited to small database.	Yes
5.	Axita Shah et al. [5]	Indian Agriculture survey database	Combination of keyword based method and natural language processing method.	53% increment	Yes
6.	Filbert Reinaldha et al. [6]	Employee database with staffing domain that uses English, Zoe is database with book domain that uses Bahasa Indonesia.	Ontology for unit conversion and Stanford dependency for natural language query conversion	Average Accuracy	Yes
7.	Azilawati Azizan et al. [7]	Google for document retrieval	Ontology and keyword based document retrieval as a use case in search engine	More accurate with ontology	Yes
8.	Manavalan et al. [8]	Biodiversity databases, GARUDA Grid project, India	Four different algorithms for different types of user queries.	Greater than 90% for the Biodiversity dataset.	Yes
9.	Sanket S.Pawar et al. [9]	IMDB and Reuters dataset (21578 files)	Extension of search engine method to Relational database search and keyword based approach.	No standardized Evaluation parameters	Yes
10.	Xuan Xuan et al. [10]	Remote sensing (RS) databases	Keyword based approach applied on remote sensing database and solves the Chinese language queries.	Development in in prototype stage	Yes
11.	Xu Yiqiu et al. [11]	Pubs database in SQLServer	NLIDBs system model and design framework based on Ontology.	87.5%	Yes
12.	Mahesh P.Gaikwad et al. [12]	Not mentioned	Approach which uses elimination method to select the tables, attributes and other values in the query.	Not mentioned	Yes
13.	Pooja A.Dhomne et al. [13]	Sample Employee Database	Pattern Matching System, Syntax Based System, Semantic Grammar System and Intermediate Representation Language are used.	Accurate to only simple queries	Yes

14.	Avinash J. Agrawal et al. [14]	Railway Database	Describes a method for semantic analysis of natural language queries for Natural Language Interface to Database (NLIDB) using domain ontology	90.33%	Yes
15.	Rukshan Alexander et al. [15]	University Database	Natural language Web interface using mapping sql template.	Accurate as per the result got from the test data.	Yes
16.	Gaganpreet Kaur et al. [16]	Not available	Used regular expression validate, filter the text in the natural language query. Constrained sequence alignment process that convert the output of one step into format required by next step.	More accurate with the dynamic regular expression method.	Yes
17.	Akshay G. Satav et al. [17]	Online applications	Broken these tokens are put in appropriate place in the SQL syntax using mapping rules	Generates the output irrespective of the database to a certain extent.	Yes
18.	K. Javubar Sathick et al. [18]	Data collected from Social media	Generating SQL query for natural language query and slightly moved to extract the data from social media websites	For a recall of 1.00 precision is 0.0014 (recall = number of relevant words retrieved /number of relevant words in sentence Precision = number of relevant words retrieved/total number of word retrieved	Yes
19.	Rongrong Zhang et al. [19]	Book database	Application of understanding the short domain question in natural language to query the data	Higher accuracy	Yes
20.	Johanna Monti et al. [20]	Not mentioned	Methodology for the development of an ontology-based Cross-Language Information Retrieval (CLIR) application and shows how it is possible to achieve the translation of Natural Language (NL) queries.	Not mentioned	Yes
21.	Mochamad Vicky Ghani Aziz et al. [21]	Tweets	Application of Natural language processing in extracting traffic information from social media data.	62%	Yes
22.	Belainine Billal et al. [22]	Twitter data	Proposed an efficient preprocessing technique for tweets.	87.6%	Yes
23.	Abdeljalil El Abdouli et al. [23]	Twitter data	Propose an approach for analyzing the data generated by Moroccan users in the social network Twitter, in order to discover the subjects that interest Moroccan society and then locate on Moroccan map the areas from where the	Not mentioned	Yes

			tweets come from.		
24.	Mohd Suhairi Md Suhaimin et al. [24]	Economic news from Facebook public pages	A method to find the sarcasm text in the social media data.	75%	Yes
25.	Li Xin, Xie Qianqian et al. [25]	Perovskite solar cells	Proposes a social awareness analysis method based on twitter data mining	Not mentioned	Yes
26.	Kaixian Yu et al. [26]	Biomedical database	Approach to extract the relationships information between protein-protein interactions over a biomedical data based on Nature Language Processing (NLP) and graph theoretic algorithm	86.5	Yes
27.	Ranjeet Devarakonda et al. [27]	Twitter, Cassandra DB, and log database	Natural Language Processing (NLP) to make the data of the Atmospheric Radiation Measurement Data Center (ADC) more accessible to the public.	Not mentioned	Yes
28.	Donia Gamal et al. [28]	Survey paper	Opinion mining using Machine learning and Natural language processing	59-90% based on the classifier	Yes

5. CONCLUSION

A systematic review of Information retrieval using Natural language processing is done and presented in this paper. Different Natural language processing techniques are used for retrieving the information from relational databases as well as social media data. Natural language questions are translated into SQL or equivalent language which the system can understand and retrieve the information from the database. Various techniques of retrieving the information from social media data which uses unstructured dataset are discussed. Tweets extracted from Twitter, facebook posts etc, are some of the common unstructured data. Most of the proposed systems are based on the keyword based. Some of them use parsing methods and different levels of language. Morphological phase, syntactic analysis, semantic analysis, intermediate language processing and finally conversion are the general steps in any of the approach. Some implementations are able to get good accuracy by making various enhancements. It is equally important to retrieve the data from social media and e-commerce data. Data storage technique is important aspect in case of unstructured data.

A deep analysis of different methods is done which includes the details of dataset used in the proposed method, approach to the problem, accuracy and an opinion whether there is a scope for improvement. Main steps involved are Tokenization, stop words removal, then linguistic analysis of the tokens by using different parsing methods and forming the queries and retrieving the result. Some of the systems are tested on a small database and cannot be extended to the larger set. Accuracy is less in some of the system due to lack of technical stuffs. Accuracy will also depend on the ontology and type of knowledge base used. More the sophisticated knowledge base, higher the accuracy. Some papers use advanced preprocessing techniques and regular expressions to increase the correctness.

After analyzing the work carried out by different researchers, there is still a lot of scope for improvement in terms of retrieval accuracy and retrieval time complexity. Accuracy can be improved by combining the different approaches. Keyword based approach will be successful only if we have large set of data specific to the database. It is not practical to have each and every possible word in the knowledge base. Hence this limitation can be overcome by combining semantic analysis with the keyword based approach. The execution time can be enhanced by storing the data dictionary using efficient data structures such as B-tree and B+ tree. This makes the searching easier by reducing keyword comparison time.

6. REFERENCES

[1] Fei L, H. V. Jagadish, “Constructing an Interactive Natural Language Interface for Relational Databases”, Proceedings of the VLDB Endowment, Vol. 8, No. 1, 2014

[2] Garima Singh, Arun Solanki, “An algorithm to transform natural language into SQL queries for relational databases”, IAEES, 3(3), pp 100-116, 2016

[3] Subhabrata Sengupta, Prasun Kanti Ghosh, Saparja Dey “Automatic SQL Query Formation from Natural Language Query”, International Journal of Computer Applications (0975 – 8887), July 2014

[4] Prabhdeep Kaur, Shruthi J, “Conversion of Natural Language Query to Sql”, International Journal of

Engineering Sciences & Emerging Technologies, Vol. 8, Issue 4, pp 208-212, January 2016.

[5] Axita Shah, Dr. Jyoti Pareek, Hemal Patel, Namrata Panchal, "NLKBIDB - Natural Language and Keyword Based Interface to Database", International Conference on Advances in Computing, Communications and Informatics (ICACCI) IEEE, pp 1569-1576, 2013

[6] Filbert Reinaldha, Tricya E. Widagdo, S.T., M.Sc., "Natural Language Interfaces to Database (NLIDB): Question Handling and Unit Conversion", IEEE, 2014

[7] Azilawati Azizan, Zainab Abu, Shahrul Azman Noah, "Query Reformulation Using Ontology and Keyword for Durian Web Search", Third International Conference on Information Retrieval and Knowledge Management, pp 94-100, 2016

[8] Manavalan, Subrata Chattopadhyay, Mangala, Prahlada Rao, Sarat Chandra Babu, Akhil Kulkarni, "Experiments on Information Retrieval Mechanisms for Distributed Biodiversity Databases Environment", IC3I, IEEE, pp 219–223, 2014

[9] Sanket S.Pawar, Abhijeet Manepatil, Aniket Kadam, Prajakta Jagtap, "Keyword Search in Information Retrieval and Relational Database System: Two Class View", ICEEOT, pp 4534-4540, 2016

[10] Xuan Xuan, Liu Jianbo, Yang Jin, "Research on the natural language querying for remote sensing databases", International Conference on Computer Science and Service System, pp 228-231, 2012

[11] Xu Yiqiu Wang Liwei Yan Shi, "The Study on Natural Language Interface of Relational Databases", 2nd Conference on Environmental Science and Information Application Technology, pp 596-599, 2010

[12] Mahesh P.Gaikwad, Natural Language Interface to Database, International Journal of Engineering and Innovative Technology (IJEIT) Vol. 2, Issue 8, pp 153-155, February 2013

[13] Pooja A.Dhomne, Sheetal R.Gajbhiye, Tejaswini S.Warambhe, Vaishali B.Bhagat, "ACCESSING DATABASE USING NLP", IJRET, Vol. 02, Issue 12, pp 589-594, December, 2013

[14] Avinash J. Agrawal, Dr. O. G. Kakde, "Semantic Analysis of Natural Language Queries Using Domain Ontology for Information Access from Database", I.J. Intelligent Systems and Applications, Vol. 12, pp 81-90, 2013

[15] Rukshan Alexander, Prashanthi Rukshan,Sinnathamby Mahesan, "Natural Language Web Interface for Database (NLWIDB)", Proceedings of the Third International Symposium, SEUSL: 6-7 July 2013

[16] Gaganpreet Kaur, "Usage Of Regular Expressions In NLP", IJRET, Vol. 03 Issue. 01, pp 168-174, January 2014

[17] Akshay G. Satav, Archana B. Ausekar, Radhika M. Bihani, Mr Abid Shaikh," A Proposed Natural Language Query Processing System", International Journal of Science and Applied Information Technology, Volume 3, No.2, pp 37-39, April 2014

[18] K. Javubar Sathick, A. Jaya, "Natural language to SQL Generation for Semantic Knowledge Extraction in Social Web Sources", Vol. 8(1), pp 01–10, January 2015

[19] Rongrong Zhang, Qingtian Zeng, Sen Feng, "Data Query Using Short Domain Question in Natural Language", IEEE, pp 351-354, 2010

[20] Johanna Monti, Mario Monteleone, Maria Pia di Buono, Federica Marano, "Natural Language Processing and Big Data An Ontology-Based Approach for Cross-Lingual Information Retrieval", IEEE, pp 725-731, 2013

[21] Mochamad Vicky Ghani Aziz, Ary Setijadi Prihatmanto, Diotra Henriyan, Rifki Wjay, "Design and Implementation of Natural Language Processing with Syntax and Semantic Analysis for Extract Traffic Conditions from Social Media Data", IEEE 5th International Conference on System Engineering and Technology, pp 43-18, Aug. 10 - 11, 2015

[22] Belainine Billal, Alexsandro Fonseca and Fatiha Sadat," Efficient Natural Language Pre-processing for Analyzing Large Data Sets", IEEE International Conference on Big Data, pp 3864- 3871, 2016

[23] Abdeljalil EL ABDOULI, Larbi HASSOUNI, Houda ANOUN, Mining Tweets of Moroccan Users using the Framework Hadoop, NLP, K-means and Basemap", IEEE, 2017

[24] Mohd Suhairi Md Suhaimin, Mohd Hanafi Ahmad Hijazi, Rayner Alfred and Frans Coenen, "Natural Language Processing Based Features for Sarcasm Detection: An Investigation Using Bilingual Social Media Texts", ICIT, pp 703-709, 2017

[25] Li Xin, Xie Qianqian1, Huang Lucheng, Yuan Zhou, "Twitter Data Mining for the Social Awareness of Emerging Technologies", Proceedings of PICMET '17, IEEE, 2017

[26] Kaixian Yu, Tingting Zhao, Peixiang Zhao, Jinfeng Zhang, "Extraction of protein-protein interactions using natural language processing based pattern matching" IEEE International Conference on Bioinformatics and Biomedicine, pp 1292- 1295, 2017

[27] Ranjeet Devarakonda (Senior Member, IEEE), Michael Giansiracusa, Jitendra Kumar, and Harold, "Social Media Based NPL System to Find and Retrieve ARM Data: Concept Paper", IEEE International Conference on Big Data, pp 4736- 4737, 2017

[28] Donia Gamal, Marco Alfonse, El-syed M. El-Horbaty, Abdel-Badeeh M.Salem, " A comparative study on Opinion Mining Algorithms of Social Media Statuses", ICICIS, pp 385- 390, 2017

[29] Marie-Catherine de Marneffe and Christopher D. Manning, "Stanford typed dependencies manual" September 2008.

[30] Valentin Ilyich Spitkovsky, "Grammar Induction and Parsing with Dependency-And-Boundary Models", December 2013

[31] http://www.nltk.org/book/ch10.html

[32] http://www.nltk.org/book/