CHAPTER 1
INTRODUCTION

1.1 OVERVIEW OF QUESTION ANSWERING SYSTEM

E-Learning is the trending technology that can provide many information resources to the user. MOOC, Moodle, Coursera, IIT - Bombay and many open sources are available for doing online courses. Learners need to clarify their doubts while doing their courses. Learners are highly depended on internet resources for clearing their doubts. But the web sources are highly flooded with relevant and irrelevant information. There are numerous search engines presented to offer wonderful information to the user. But the user should explore numerous pages to navigate to discover their significant information. Hence they are in need of authenticated answers for their doubtful queries. Question Answering System (QAS) supports the query posted by the user with Information Retrieval (IR). Question Answering System is defined as answering the user query by retrieving the answers from the web source with the support of Natural Language Processing (NLP) and Information Retrieval.

There is a lot of Question Answering System available and it has been applied in various fields such as social media, recommendation systems, natural sciences, medical sciences, e-commerce, religion, and education system which help to provide answers related to the queries posted by the user. The wide use of QAS in many fields indicates that it contributes to improving the quality of life. Currently, various QAS supports the users for clearing their doubts. But it is the social responsibility of the researcher to provide the most relevant and authenticated answer for their posted query. In any QAS, the user has to post the question and it is analyzed based on the keyword and the suitable answer is retrieved from the pool or repository using the keyword.
Question Answering System is classified as Open Domain QAS in which questions are posted in any domain and answers are retrieved by the learners whereas closed domain QAS the questions are posted in a particular domain and the user will get the suitable relevant answer. Open domain Question Answering Systems like Yahoo Answers.com, Askme.com, About.com, and stackoverflow.com are serving the users by providing different answers to their queries. It is the duty of any QAS to provide the most relevant answer to the question posted by the learner for clarifying their doubts.

In the existing QAS, the answers are retrieved in many ways such as pattern matching, template matching, passage retrieval and structural based retrieval. But the retrieved answers may not be suitable for the query posted by the learner and may not satisfy the user’s need also. Hence exact retrieval of answer and validation of answer is mandatory for clarifying the learner’s doubt.

In passage retrieval, the keywords in the query have been identified and retrieved based on the key feature. There is no guarantee that the retrieved answer may satisfy the user’s doubt. The retrieval may have happened based on the static mapping. It is required that the semantic analysis of question and answers has been done before handover the answer to the user. In some QAS, the question and their answer will be retrieved based on FAQ in order to save the time. But the FAQ (Frequently Asked Questions) may not showcase the question what the user posted or the user may not retrieve the required answer.

In pattern matching approach, the answers were searched from the corpus based on the frequently asked question types. These QA systems which support pattern match technique will understand the pattern of text from the question which is posted by the user for retrieving answers. Pattern
matching approach can be classified as the surface text patterns and templates matching approach. Surface Pattern approach extracts answers to the question from the surface structure which consists of a list of surface patterns. The answer is mapped to the question which is relying on the surface patterns. The similar pattern has to be identifying while mapping the question with their answer using a regular expression.

A template based approach is a type of pattern matching approach which is having preformatted patterns called templates for questions. The templates are built which contains the number of templates ensuring that it confer the answer for their question posted by the user. For every question type, the template has to be built based on that the answer is mapped to the question. It also consists of entity slots used for the missed questions which are not available in the repository. After retrieving answer for the posted question, the template is filled up with the question as well as the retrieved answer. This response returned by question with its answer is returned to the user. But the most important challenges in QAS is to finding the exact answer to the user's question.

In many QAS, the answers are collected from different sources and also the answers are retrieved from the repository. This repository needs to be updated frequently to get the relevant answers with the help of dynamic sources. But most of the QAS mostly rely on static data. Many QAS is not involving with collaborative learning. In collaborative learning, the users got profited when the learner uncovered to different perspective from the other learners with varied backgrounds and gives the most excellent interactive teaching-learning methodology where the interaction has taken place between the learners and also with the experts. It helps to involve a number of people to give their answers and also the answers are authenticated by the expert. Even though the data are collected from the different sources, there is a need for an expert to verify the answer.
Mapping of the question with their corresponding answers plays a vital role in any QAS. The existing system mapping was done with the keyword similarity which lagging in the semanticness of the answer. So the answer needs to be validated with the help of reasoner. Many QAS system is not supporting the rules which help to give reasoning the answer and also it is lacking in ordering the retrieved answers.

The reasoning based question answering system is relying on the KBQA for providing the most relevant answer to the question posted by the user. Knowledge Based Question Answering System (KBQA) solves the issues of passage retrieval where the syntactic and semantic nature of the question and the answers are analyzed with the support of ontology. It gives the conceptual meaning of the keyword and it provides more meaningful search results by evaluating and understanding the keyword and finding the most relevant answer from the web, database, collaborative forum and the expert’s answers. Ontology is used to improve the semantics of the query.

The answers are retrieved statically which poses the need for dynamic retrieval of answers if they are not in the repository. Collaborative learning helps for dynamic retrieval of answers which is incorporated in KBQA. It involves multiple users to give their answers. The answers are collected from various resources like collaborative forum, web, and repository and it has to be analyzed syntactically and semantically. After that, the requested question is mapped with relevant answers using concept mapping and rule based mapping. LPLSA is integrated with description logic, SWRL, and SPARQL Queries to provide more semanticness to the query posted.

The correctness of the mapped answer is verified with the reasoner which is used to derive the information from the knowledge base with the help of inference engines. The reasoner of the KBQA system is used to
validate and find the correctness of the question with its mapped answers. The KBQA system uses the default reasoner called pallet which supports to find the inconsistent ontology at the time of constructing the ontology and Fuzzy reasoner is applied when the KBQA system uses the collaborative forum where the answers are collected from the various dynamic resources and the correctness is checked with fuzzy reasoner whereas Case Based Reasoning helps to get the relevant answers when the user is trying to retrieve the answers based on their previous questions.

As a part of KBQA system, Collaborative Learning supports for answer mapping. In Collaborative Learning, the users have the opportunity to discuss with the peers, present and share their ideas, exchange their thoughts and question about other conceptual frameworks. The posted question can be answered by other users after their brainstorming session and the answers are analyzed and all the answers are prioritized based on voting in KBQA. At that point, it is passed to the experts who are authorized to check the correctness and significance of the answers to the question. Finally, the ranking process helps to give the most relevant answer to the user.

1.1.1 History of Question Answering System

![Figure 1.1: Evaluation of Question Answering System](image-url)
Figure 1.1 shows the evaluation of question answering system. The era of NLP generally started in the year of 1950s by Alan Turing and the QAS has become the part of NLP. The famous QAS are developed in the year of 1960s namely BASEBALL and LUNAR. BASEBALL is the QAS which is able to answer questions approximately for the US baseball association over a period of one year whereas LUNAR is replied the questions almost the geological investigation of rocks returned by the Apollo moon missions. LUNAR was illustrated at a lunar science convention in 1971 and it was able to give reply about 90% of the questions posted by people untrained on the system.

The era of Text comprehension QAS projects started in the year of 1970s and 1980s and the first system was the Unix Consultant (UC) which was answered questions about Unix operating system. LILOG was another QAS that operated about the tourism information of a German city. All these QAS are the Closed Domain QAS.

The evaluation of Open Domain QAS such as Yahoo! answers were started on June 28, 2005. Initially, Yahoo! Answers named as Yahoo! Q & A which is a community based QAS. The developers are Jerry Yang and David Filo who created the Yahoo website in the year of January 1994. The users are allowed to ask question as well as to submit answers for other user's question. It encourages the users to earn points based on their participation. It follows Naver's Knowledge iN and available in 12 languages. Questions are initially open to answers for four days and allowed to answer within minimum of one hour to the maximum of up to eight days. Users who provide the best answer is selected as the top contributor. One user can be the top contributor to three community.

Answers.com was developed by the entrepreneurs Bill Gross and Henrik Jones in the year of 1996 which includes WikiAnswers,
ReferenceAnswers, and VideoAnswers. It supports to answer the questions in different languages. The website domain was acquired from GuruNet and initially it displayed the information through the downloadable software product. In 2006, answers.com released a trivia game called blufr and the revised version was released in the year of 2009 which combined wiki-style contributions.

Another QAS Ask.com which is an open domain QAS originally known as Ask Jeeves was developed in the year of 1996 by Garrett and David Warthen. It focused to answer about e-business and it is currently owned by InterActiveCorp (IAC). In the year of 2010, it released beta Q&A service and launched its mobile app for iPhone.

Siri is an intelligent personal assistant developed by Apple Inc and integrated into iPhone 4S in the year of 2011. It uses speech recognition engine provided by Nuance Communications and collects the user’s query in the form of voice in a natural language user interface for answering the questions.

IBM Watson was developed in the year of 2011 and is running a softer called Deep QA which is understanding the tricky question and finding the best answer to that question. It is able to give the most relevant answer for the unstructured data. Watson runs on a cluster of Power seven fifty computers holding ninety servers and 2889 processor. Watson was able to answer any question with the support of resources, including World Book Encyclopaedia, Wikipedia, and books from Project Gutenberg.

1.1.2 Types of Question Answering System

The Question Answering Systems can be classified as Open Domain QA System and Closed Domain QA System. Open domain QA System
deals with any domain of questions. It can discuss any topic under any domain and there are no restrictions on the open domain question answering system. The data can be collected from the various resources for answering the question posted by the users. The data sources acts as the repository and the open question answering system rely on the internet resources. General ontology can be used for constructing the open domain question answering system whereas closed domain QA system deals with questions under a specific domain.

For constructing the closed domain question answering, the domain has to be fixed. Based on the domain, the repository is created. When the user is posting the question, the answer is searched from the repository. Ontology is created under the specific domain and only limited types of questions can be answered. The questions may be descriptive rather than procedural.

![Diagram of Question Answering System](image)

**Figure 1.2: Types of Question Answering System**

In Factoid Question Answering Systems, the questions and their answers are focused either on syntactic and/or semantic information. Based
on the interrogative words the question can be answered by the answer keyword organization or person names. It consists of three phases such as Question Processing (QP) phase, which identifies the type of the question, followed by Passage Retrieval (PR) phase, which extracts the number of relevant answers from the repository, and finally Answer Extraction (AE) phase, which ranks exact answers from the previously retrieved answer pool which is stored in the repository and the relevant answer is given back to the user.

The QP component phase is the first component of factoid QAS, which identifies the type of the input questions posted by the user and mapping them into the taxonomy. There are two level taxonomy used which consists of 6 question types and 53 subtypes. In the PR phase, the process is divided into two steps and in the first step, the question words sorted in descending order based on their priority. In the second step, the set of keywords are identified and used for retrieval of answers. The Answer Extraction (AE) component identifies relevant answers from the passage repository set and extract the answers to the user's question.

Web based question answering systems are highly relying on the search engines to retrieve webpage’s that contains answers to the questions. Most of the Web based QAS are the open domain while some of them work for domain oriented also. The web based QA systems such as MULDER, NSIR, ANSWERBUS falls into the category of domain independent QA systems, while START and Katz.B are referred to as domain specific QA systems. The Web Based QAS handle wh questions such as “who has invented java”? This QAS provides answers with the support of text documents, hyperlinks, XML documents, Wikipedia and search engines like Bing, Google, Yahoo, Alta Vista etc... Based on the architecture of Web QAS, three common levels are used for providing the answers for the user’s question.
The first level is the question classification level which classifies the user’s question into one of the question types. From the question classification, the answers are processed using the answer extraction level. The answer extraction level extracts the correct plausible answers to the questions. The answers are collected from the various web pages and it is stored in the repository. The answers also searched in the search engine which may give a number of answers for the single user query. So it is mandatory to identify the relevant answer which is carried out in the answer selection level. Among the plausible answers collected from the various sources, ranking techniques are applied to find the best answers based on their similarity score and the answer which is ranked in the top place is considered as the final answer.

One of the popular open domain web based QAS is LAMP which makes use of snippet tolerant property to calculate ranking percentages based on which responses are produced by Dialogue based QA systems. WEBCOOP is a cooperative type of web based QAS which integrates knowledge representation and reasoning approaches to generate the cooperative responses. Most of the Web based QA systems perform a better effort to produce correct answers. Though these systems are most capable to handle wh-type of questions but lack to produce accurate answers, instead these systems retrieve the relevant passages that contain keywords to extract the answer from the knowledge base which requires an additional process to obtain the exact answer. This QA system fails to provide answers to temporal based queries.

In IR / IE Based QAS, the user question returns a set of top priority documents or passages as the reply to the question. Information Extraction (IE) system uses paring to parse the question to pre-process the user’s query. IE systems depend on Named Entity Tagging (NET), Template Element (TE), Template Relation (TR), Correlated Element (CE), and
General Element (GE) for processing the request. IE systems framework consists of different levels like Level 1, Level 2, and Level 3. Level 1 acts as a NE tagger which used for pre-processing the question based on the interrogative words such as who, when, what, where etc.,

The level 2 handles the named entity tagging along with their adjective such as how far, how long, how much etc., Level 3 plays the major role by processing the whole sentence with the support of General Element. For example “Who is the father of computer?” is passed onto the levels mentioned above ASKING POINT is Person (Noun) KEYWORDS such as father and computer are retrieved. The framework of IE systems includes of two phases, such as question processor and text processor.

- **Question processor** phase is used to get the input from the user and generates the keyword for the question which supports to match the answer from the repository.

- **Text Processor** retrieves keywords from the text with the support of named entity recognizer to generate accurate answers.

The IR Based systems like AskJeeves, LaSiE system uses some basic phases such as Tokenizer, Sentence splitter, Parse process, Name matcher, Discourse Interpreter to perform the text analysis. It depends on the repository which needs an expansion of CE and GE components to handle yes/no types of questions within the text. These systems can answer only wh-type of questions but other than wh-type of questions such as “How can I get the answer to the user’s question?” remains unanswered.

Restricted Domain Question Answering systems is a closed domain QAS which requires a linguistic support to understand the natural language text in order to answer the questions correctly. It improves the accuracy of QAS by restricting the domain of questions, the type of the questions and the
size of the repository which is directed in the development of Restricted Domain Question Answering System (RDQA). RDQA have specific characteristics like providing accurate answer for the question and reducing the level of redundancy. It overcomes the difficulties faces in open domain QAS and improves better accuracy. LUNAR is the example of RDQA which allows answering geologist analysis of rocks. BASEBALL is another restricted domain QA system, which can only answer about one season’s Baseball data. These early QAS have incorporated a huge amount of domain knowledge in databases.

RDQAS doesn’t focus on language understanding it focuses on a specific set of domain rules. The RDQA systems use IE engines which consist of web crawlers and Wrappers, WebCrawler is used to select set of extraction rules that can extract domain information, while Wrappers used to retrieve relevant domain oriented WebPages which contains answers. RDQA first analyzes questions and with the support of Structured Query Language (SQL), the queries are translated to obtain results by retrieving data from the database.

Domain Oriented systems make use of domain oriented knowledge bases, domain servers, information systems, parsers etc. Knowledge Acquisition and Access Systems such as KAAS uses IR systems to retrieve relevant passages that are processed by NLP. Domain oriented QA systems contribute great effort to achieve accuracy from the data which is retrieved by the information retrieval. The system requires situated evaluations while comparing with that of open domain QA systems. This type of QA systems needs a domain classifier to handle different domains which is one of the difficult tasks to achieve. There are different approaches used to process the questions and the answers. The first most important approach is Parsing, Tokenization help to pre-process the user’s question.
Next approach is the statistical approach which is implemented in online text processing system. Statistical approaches are not depending of SQL queries. It can be processed in natural language form. One of the drawbacks of the statistical approach is that it process each term independently and unable to identify linguistic features for a combination word or phrase. Different statistical techniques are used in every stage of the QAS. Some of the techniques are Maximum entropy models, support vector machine (SVM) classifiers, Bayesian classifiers. One of the example for the statistical method was IBM’s statistical QA system. It uses maximum entropy model for question and answer processing based on various N-gram features.

In pattern matching approach, the answers were searched from the corpus based on the frequently asked question types. These QA systems which support pattern match technique will understand the pattern of text from the question which is posted by the user for retrieving answers. Pattern matching approach can be classified as the surface text patterns and templates matching approach. Surface Pattern approach extracts answers to the question from the surface structure which consists of a list of surface patterns. The answer is mapped to the question which is relying on the surface patterns. The similar pattern has to be identifying while mapping the question with their answer using a regular expression.

A template based approach is a type of pattern matching approach which is having preformatted patterns called templates for questions. The templates are built which contains the number of templates ensuring that it confer the answer for their question posted by the user. For every question type, the template has to be built based on that the answer is mapped to the question. It also consists of entity slots used for the missed questions which are not available in the repository. After retrieving answer for the posted question, the template is filled up with the question as well as the retrieved answer. This response returned by question with its answer is returned to the
user. But the most important challenges in QAS is to finding the exact answer to the user's question.

Semantic analysis approach supports for term definition and query expansion processing to process users queries. This technique is able to recognize the possible meanings of the questions from words that used in the question. However, this technique has drawbacks, i.e. assuming that the source texts are expressed in natural sentences and ontologies-semantic only cover a particular domain of knowledge. Hybrid Approach is a combination of linguistic, statistical, semantic, and rule based pattern matching technique to analyze query.

1.1.3 Need for Question Answering System

The various needs have been identified to develop a new question answering system. The first and most important need is to identify the question classes. There are many types of question classes need to be identified and the question classes can be classified into descriptive class, enumeration class, Yes/No class, choice class, define class etc., These question taxonomy of the QAS needs to be identified. Next, consideration is about the question processing. Some questions may be processed by syntactically and some may be processed semantically. And also some questions may be processed based on interrogative words and some may be based on assertive. Hence the processing of question must be identified.

Consecutively, the collection of answers from the data source needs to be identified. The data source may be static or dynamic. The answers can be collected from the various data source to form the repository of the QAS. The collected answers need to be processed and evaluated for giving the most relevant answer to the user. If the answer is not available in the repository, then it is not possible to obtain the answer for the question. Any QAS, the
answer has to be extracted from the different data sources and the complexity of the question as well as the answer need to be analyzed. The context of the question has to be maintained and based on that the answers have to be retrieved from the data sources.

The answers are collected from the various sources and comprehended to form a single answer. This may be the toughest task because, at the time of computing the answers, it may lead to produce the incorrect answer. When evaluating the answers, the contextual meanings of the answers are needed to be considered. From the context of the question, the complexity of question can be analyzed to understand the question. Most of the QAS system, navigate the pages for getting the context of the answers and the navigation process cannot give most suitable answers. Hence, it is a need to identify the context of the answer based on the question asked by the user.

The correctness of the answers needs to be checked with the support of reasoning. Reasoning capabilities must be added to upgrade the QAS to improve knowledge base. Based on the requirements different kind of reasoning must be identified and applied to validate the correctness of the QAS. Some of the reasoners are Common sense reasoning, Forward chaining mechanism, backward chaining mechanism, case based reasoning etc.,

The major important factor needs to be considered while designing a QAS is collaborative learning. In collaborative learning, different ideas are collected from the various learners and the ideas are shared to find an answer to the question. The learners are categorized into different level based ability. To identify the level of user ability, QAS need to collect the learner's information. It makes the necessity of developing user profiling in QAS and also the necessity of interactive QAS is required which helps to
interact with all level of users. Interactive QAS supports to share their ideas immediately and the user will give their acknowledgment immediately.

Most of the QAS systems are developed in the English language which creates the need of other languages in QAS. Multilingual QAS supports to interact with the user in their language which makes the user to feel comfortable to communicate with the system.

1.2 MOTIVATION

The method of answer retrieval is still based on the syntax, which leads to the lack of semantics in question and answers analysis. The answer is stored and retrieved statically in many question answering system, which leads to developing a dynamic knowledge based question answering system. Because of the static data source, the mapping question with their answer is done in a structured manner, which leads to providing a QAS with rule based, and description logic based system. Many QAS system is not involving the expert’s knowledge and the different user’s group may not be an expert in the domain, which leads to developing a collaborative question answering system. The Correctness of the answers has to be checked before the answer is given back to the users. Many question answering systems are lacking in reasoning.

1.3 AIM AND OBJECTIVE

The objective of the research is to provide the authenticated answer to the query posted by the user with the support of ontology, web rule language, description logic, and reasoning. Reasoning Based Question Answering System is relying on Knowledge Based Question Answering System (KBQA) for answering the question posted by the user. The key objectives of the reasoning based question answering system are given below:
• To analyze the questions and answers both semantically and syntactically with the support of KBQA ontology. Ontology supports for semantic analysis of questions and answers by preserving the semantic concepts and the relations of the questions and the answers.

• To map the retrieved answers to their suitable questions posted by the users where the answers are retrieved from the repository and web sources.

• To validate and improve the correctness of the answers based on the semantic context of the answers. Case Based Reasoning (CBR) and Fuzzy Reasoning are supporting the KBQA System. To improve the performance and correctness of the system, the rules are derived and incorporated with the KBQA system.

• To provide an ordered list of answers based on the semantic closure to the query.

• LP-LSA technique is derived for mapping the question with their answers and Descriptive logic, SWRL Rules is integrated with LP-LSA to strengthen the KBQA system.

• With the support of ONTO-KBQA ontology, the semanticness of the KBQA system is preserved in every aspect.

• The collaborative forum is tied up with the support of fuzzy reasoning which leads to the interactive QAS. It also involves more learners to provide the answers. The reasoning is applied in every aspect to validate the correctness of the mapped question with their answers.
1.4 ORGANIZATION OF CHAPTERS

The contributions made in this thesis are organized into eight chapters and is explained as follows:

The present chapter provides an introduction to question answering system. The history and types of QA are discussed in a detailed manner. Then the various issues in the QA systems are highlighted and the proposed research objective is explained.

Chapter 2 deals the background and detailed literature review of related work in the question answering system. A detailed comparison made on the works of different authors has been discussed in detail.

Chapter 3 focuses the semantic analyzer for questions and answers which deals the question and answer evaluation. It provides the framework of KBQA System and their four phases are explained in detail. The first part of this chapter explains the construction of ontology which is the base of KBQA system. Then the question processing methods and the answer evaluation procedures are discussed with sample data set.

Chapter 4 presents the rule based mapping of questions and answers and also discussed the concept mapping. The LP-LSA which is the method used in KBQA is analyzed and proved. The description logic, SPARQL and the SWRL rules used for the inference is also discussed. Finally, the chapter ends with the brief description of the reasoning.

Chapter 5 shows the semantic reasoner for questions and answers which deals with the various reasoning technique like case based reasoning, fuzzy reasoning, and Pellet reasoner. This chapter also discusses the send-
a-problem method of collaborative learning and the second half of the chapter briefly shows the comparative results.

Chapter 6 gives the ranking of answers which ranks the most relevant answer to the user gives the detailed results of the KBQA System.

Finally, Chapter 7 concludes the research work by giving brief glimpses into the future direction of the research.