CHAPTER 2
LITERATURE OVERVIEW

This chapter provides a detailed survey of the investigations about various studies, models, methodologies and evaluation methods of Question Answering System. The different aspect of the QAS was discussed by referring the various authors. The summarization lists the definitions and discussions on the basic information related to Question Answering System, which is important for designing any successful QA system which guides the developer for developing any QAS.

2.1 QUESTION ANSWERING SYSTEM

Question Answering System of the first generation appeared before 1965, had a limitation on a linguistic model. The second generation of QAS preceded by ELIZA computer program was built by J. Weizenbaum (1966) proposed a QAS called ELIZA is a program which takes the natural language conversation between a man and the computer. It is operating with the MAC time sharing system and it is developed by MIT [1]. Here the input sentence is analyzed and the keywords were used to map the answer from the input sentence. It was done based on the decomposition rules.

W. A. Woods (1973) developed the Lunar Sciences Natural Language Information System (LUNAR) was a research which deals the geological analysis of rocks returned by the Apollo moon missions [3]. It was answering the LUNAR system and demonstrated at a lunar science convention. It was able to answer 90% of the questions in its domain posed by untrained people. It underpins the man-machine communication issues by adjusting the machine to the traditions of standard natural English instead of requiring the man to adapt to the machine.
D. L. Waltz (1978) proposed a question answering system which is able to answer about the air aircraft flight and planes with the help of large relational database [4]. The language processing portion of the system uses a number of networks which matches traces with a specific meaning for judging the questions and cater the answer for the user. A story understanding system (QUALM) is described by Lehnert (1981). It works by asking the questions about simple and lengthy paragraph stories. QUALM system framework includes question analysis phase that links with its question type. Based on the analysis of question type, the answer is processed and retrieved from the repository.

Salton and McGill (1983) described a QAS that can able to provide the answers for the posted questions directly. Kupiec (1993) employed a QAS which deals with the WH-questions. The question was processed based on the interrogative words for retrieving the answers by the system [5]. In recent 90’s, question answering achieved a great progress due to the Text Retrieval Conference (TREC), which has consisted of a textual question answering session since 1999, with a wide range of research groups participating, both from industry and academia.

S. Tellex et al (2003) implemented a QAS framework named Pauchok which supports passage retrieval for QAS to retrieving the answer from the answer pool [11]. Passage retrieval phase process the set of answers and provide the ranked lists of passages with respect to the question. For few questions, this method does not provide any correct answers in the ranked list of top twenty answers.

R. Soricut et al (2006) described the framework tritus in which the answers were retri by existing information sources and the repository. The questions are transformed into new queries to provide the answers and it depends on the existing information retrieval systems [22].
consists of three phases such as automatically learning phrase, candidate generation and the candidate evaluation. In automatically learning phrase, the question classification and its different types were analysed. In automatically generating candidate phase, the question is transformed into a training set of question/answer pairs and automatically evaluating phase, the candidate answers were retrieved from the target information retrieval systems such as search engines. Reranking is performed on the retrieved answers and reshuffled at run-time.

Vanessa Lopez et al. (2006) discussed about Power Aqua in the research paper titled “Question Answering on the Real Semantic Web”. It was dealing with the semantic data of Power Aqua balances and giving results with the support of ontologies [24]. It translated the user query into answers by semantically with the deep terminology. Asserted model was created for the terms in the query and the answers were from various knowledge sources. The multiple sources and domains were combined to provide the best result for the user’s query.

Abdullah M. Moussa (2011) was developed QASYO[5] which is a sentence level QAS. It uses YAGO ontology which is a light weight ontology and works on natural language processing, ontologies [43] and information retrieval technologies. It consists of four phases such as question classifier, linguistic component, query generator and query processor. It used triple based data model and binary relational model. The relationship of the candidate answers has been identified with the support of WR table and Verb table and it used parse tree construction.

Kathrin Dentler et al. (2011) provides a survey and a comparison of the semantic web reasoners which is used to identify the large ontologies. The ontologies were expressed in the tractable OWL 2 EL profile [46]. The correctness of the ontologies were improved by the reasoners and they were
used several dimensions such as reasoning characteristics, practical usability. The performance indicators were evaluated by the classification, concept satisfiability, subsumption checking and consistency checking performance.

Anni-Yasmin Turhan (2011) discussed with the description logic and the inferred model [47]. The rules were defined and inferences were made based on the semantics of Description Language. These inferences have been investigated with number of description rules and expressivity. The expressivity of a DL is determined by describing the concepts and roles of the semantic data. He also discussed T-box and A-box knowledge bases and tested with different reasoning techniques.

Fernando et al (2012) described about DELOREAN, which helps to convert the fuzzy rough ontology into RDF ontology languages [54]. With the support of the parser module, the fuzzy rough ontology which is having the input physical URI and translated into a fuzzy KB. The reduction module, the fuzzy rough ontology is reduced into an equivalent non-fuzzy ontology. The inference module is used communicated with a non-fuzzy ontology reasoner in order to check the correctness of the ontology. A DL reasoner is used to reason with the resulting ontology with its all possible concepts, rules, relationship and axioms.

Dang Truong Son (2012) proposed a QAS for the Vietnamese language in which the question from the user is converted from natural language to Vietnamese language. It is accomplished by understanding the sentence with its synonyms and trimming the stemming words of the sentences [55]. The language modeling technique is used to store the user's question into the repository. Few techniques such as Part of Speech, TF-IDF methods are used to process the question as well the answer.
Athira P.M et al (2013) discussed a QAS, in the question is processed for extracting answers in natural language [58]. IR systems the relevant documents are collected by the matching words and present the set of documents that relate to the user’s query. But the information may not exactly specify the correct answer for the user’s question. The documents are retrieved by matching the keywords from the question with a set of index terms to identify the answers. In contrast, IE systems are giving most relevant answer for the question posted by the user. It also gives better accuracy than IR systems. It uses template pattern matching to map the question with their relevant answers.

Yong-Bin Kang et al(2014) developed a predictive model of QAS using regression techniques. It takes ontology metrics features as the backbone of the system [63]. They used accurate regression models and Performance hotspot identification model as a predictive model. A regression model uses six reasoners through 10-fold cross-validation which provide highly accurate results. The models are demonstrated by R2 and RMSE values between the training sets and the test sets. Performance hotspot identification is adopted to exhibit the strengths of the regression models.

G.Suresh Kumar(2014) has proposed a QAS which handles unstructured text. It extracts the concept relation from the unstructured text using Navie Bayes Classifier[64]. The concept mapping uses the Hand coded dependency parsing pattern rules and binary decision tree based rules to seed the relationship. The RDF triple set is used to extract feature of the question and the answer which supports to map the question and with its relevant answer.

Rupali R. Khune (2015) offered a mapping framework for the Multiagent ontology which is having heterogeneous data in web and assigned the semantic relation using semantic relation interpreter which
improves response time [67]. Multiagent system mapping system was developed to map the question with their answer.

T. Atapattu et al(2015) utilized the framework for extracting question-specific concept maps (QSCMap) and proposed a method that returns answers as a concept map, which further encourages meaningful learning and knowledge organization[68]. Here the Question type will be converted to triple form and they have taken Descriptive and Comparison type of question to analyze. Triple Enhancer is used to find the synonyms of the word to find the keyword from the domain model. Map merge method is used to automatically combine the concept maps of related topics by reading the CXL (Concept Map Extensible Language) file.

N. Suryana et al (2017) analyzed techniques used in question analysis, document processing and answer extraction [96]. Each and every phase of question answering system modules and their sub modules processing techniques were analyzed with their input and their outcomes based on linguistic approach, statistical approach, semantic approach and hybrid approach.

Fandy Setyo Utomo et al(2017) provided the review of Question Analysis, Document Processing, and Answer Extraction Techniques [97]. The different techniques of question answering systems were discussed based on various approaches like linguistic approach, statistical approach semantic approach, rule based approach and hybrid approach. The input and the output of every phase were discussed in a clear manner.

Sudip Mittal et al(2017) framed a question answering system for cloud computing services [98] that can be used to analyze and obtain information from the Service Level Agreements (SLA). It automatically extracts the knowledge from the SLA. The information was stored in the form of RDF
(Resource Description Framework) and it was retrieved with the help of SPARQL (Simple Protocol And RDF Query Language) Query.

Waheeb Ahmed et al (2017) developed an automatic web based question answering system for E-Learning [99]. It was detecting the question type and providing only the blocks or phases of data rather than giving full documents or hyperlinks. They were dealing the factoid question and used Support Vector Machine (SVM) for classifying the questions. It depends on the Google search engine for collecting the answers and proving the answer for the question based on the google search results. The documents were ranked with the support of n-gram and the TF-IDF (Term Frequency- Inverse Document Frequency) methods. The question and the answers were analyzed syntactically than semantically. Hence the system depends only on the keywords and their similarity measure which lacks in semanticness of the system.

2.2 SEMANTIC REASONING

Song Jun-feng et al (2005) described ontology-based QAS for the Semantic Web [21]. It uses OWL Lite as standard ontology language, which helps to construct the ontology and handles the complexity of reasoning problems. Ontology construction is also uses the integrated domain ontologies. The Meta data of web content is integrated with the ontology to mark up the Web content and it provides the semantic nature of the question as well as the answers. These semantic mark-ups are acts as the semantic index terms for extraction answers for the question posted by the users. This model ensures that it provides the most relevant and highly accurate answers for the user’s question.

Victoria Uren et al (2006) designed AquaLog QAS which is an ontology-driven QAS for the Semantic Web [24]. AquaLog has been
designed for the semantic web languages and it was integrated with different ontologies. The different ontology servers are integrated easily to provide the knowledge representation platforms.

Qinglin Guo et al (2008) proposed a QAS framework for Chinese Natural Language based on the ontology. It is a human-machine interface implemented with the support of NL parser which is used for the pre-processing of the question as well as the answer. The documents were crawled from the web and it was pre-processed by web pages crawling, HTML format filtering, segmentation, and Tagging. The extracted documents have been parsed by the NL parser to identify the relationship between the terms of the document. It uses pattern matching techniques to the research area of reading comprehension. Ontology is used to maintain the semantics of the terms in the document.

Elham S. Khorasani et al (2009) framed a Computing with Words (CW) based QAS with Reasoning Methodology. Computing with Words (CW), proposed by Zadeh [41], which offers a mathematical tool to represent and reason the question and answers with sensitive information. The correctness of the CW-based QAS is validated with the reasoning methodology. This methodology takes the question as input, extracts the answer from the knowledge base. The query is in form of generalized constraints and the query is converted into a new tree structure, called as Constraint Propagation Tree (CPT). CPT helps to identify the most relevant answer to the question posted by the user and allows to give the most accurate answer by establishing an information-seeking dialog with the user.

Paloma Moreda et al (2011) constructed a framework for Named entity question answering system which focuses on the semantic information on the question and answers posted in the system [48]. It was also considered the common noun for the pre processing of the question as well as the answers.
The pre processing was also done with the support of Semantic Role Labeling (SRL) in which rules were written to map the question with their answer. The pattern matching techniques were applied with the support of wordnet to process the answers. The system supports only on the wh questions and the semantic information system provides more precision and correctness of the answers.

Erfan Najmi et al (2013) developed an intelligent semantic question answering system which supports ontology [59]. The questions were collected from the user and it was converted into Resource Description Framework (RDF) to find the required answer. It was designed to give the answer in a single word which was lacking to provide the detailed answer to the question. It was handling the wh questions of factoid question type and supported to handle true or false question type. Parser and N-grams were used to classify and identify the keyword from the question. SPARQL Query was applied to retrieve the answer to the question.

Amir Zidi et al (2013) described a framework for ontology-based question answering system [55]. This research focused on the detection of semantic information retrieved from web sources and it is mapped with the support of knowledge base. It consists of three phases such as the knowledge phase, the indexing phase and the retrieval phase. Keyword-based semantic retrieval system is used to retrieve the question for the answer by using the domain ontology.

R.suganyakala et al (2013) described Movie related information retrieval using ontology based semantic search [60]. A domain-specific ontology is constructed for movies which act as a repository of the proposed model. The semantic nature of the system provides more accurate results for the user’s query. Ontologies are used to retrieve the knowledge sources from structured data sources.
Gagan Deepsingh Narula et al (2014) described a framework for Improving statistical Multimedia Information Retrieval (MIR) model by using ontology [62]. The multimedia content cannot be retrieved easily from the massive collection of documents. It is highly complex because the multimedia documents may include various data types such as text, images and video documents. The proposed model uses ontology and different statistical IR approaches to extract the multimedia content from the repository.

Poonam Tanwar et al (2014) created the architecture of CNLP AIDE and discussed the reasoning allotted in their research work of An Effective Reasoning Algorithm for Question Answering System [61]. The knowledge representation system acquired the knowledge collected from the sources and processed with the CNLP AIDE system. The knowledge was stored as a story and with the support of reasoning system, the new knowledge was inferred. It consists of four modules, question-answering system, document processing, Language-to-Logic (L2L), Search Engine, and Answer Providing Passages. The reasoning was also applied to the IQAS system and it used forward chaining algorithm and it was tested in declarative and procedural knowledge.

Syarilla I et al (2015) has developed a framework for the semantic question analysis model for question answering system which takes the natural language question from the user and the answers were extracted from the knowledge base [69]. The semantic analysis question model uses the Question Answering system with User modeling and Relevance Feedback (QAUF). This model considers four lexical elements which are the focus, head focus, modifier term and modified term. The system was analyzing the natural language question with the support of ontology and the keywords are extracted by TF-IDF. It was giving better result percentage in F-measure and precision.
Majid Latifi et al (2017) proposed a framework of ScoQAS: A Semantic-based Closed and Open Domain Question Answering System where the major types of question answering were discussed [100]. The closed domain question answering system depends on ontology whereas the open domain question answering depends on Linked Open Data (LOD). It uses tuple pattern for question classification where the question was analyzed with tokenization, POS tagging, and NER. But it depends on the syntactic information of the question. Rules were framed for a set of conditions and their actions. QGraph (Question Graph) was used to identify the semantic relationship of the question and the search term was extracted with the help of SPARQL Query. It was lacking to show the statistical measure to identify the accuracy of the system.

Simone Scannapieco et al (2017) was developed the open question answering system called Shoo the Spectre of Ignorance with QA2SPR [101]. An Open Domain Question Answering Architecture with Semantic Prioritisation of Roles which used the technique Prioritised Semantic Role Labelling (PSRL). PSRL uses the frame semantics theory and semantic frame representations to identify the meaning of words through schematic representations of the situations that characterize human experience, each constituted by a group of participants in the situation, or frame elements (FEs), and describe the possible syntactic realizations of the FEs for every word. Usually, the information necessary for the individuation of semantic frames is gathered by annotating corpus sentences in a specific language with FEs (semantic roles) and syntactic information. Syntactic structure matching has been applied to candidate passage retrieval and answer extraction. PSRL is a novel semantic frame-like approach for Italian logical complement analysis based on Schank verb theory.

S. Jayalakshmi et al (2017) created a framework for Automated Question Answering System Using Ontology and Semantic Role which
presents the Web and semantic Knowledge Driven automatic question answering system (WAD) [102]. Named entity and entity linking methods were applied in preprocessing of questions. Semantic of question and answers were maintained with the support of ontology and wordnet. The ranking was applied to the selected answers for the question based on the similarity of the keywords and conditional probability model was also applied to enhance the ranking process. It was working only for the factoid question and their results were discussed.

Daniel Khashabi et al (2018) proposed a question answering system in Question Answering as Global Reasoning over Semantic Abstractions which is able to answer multiple choice questions [103]. The system SEMANTICILP, reasons over a wide range of semantic abstractions of the text, which are derived using off-the-shelf, general-purpose, pre-trained natural language modules such as semantic role labelers, coreference resolvers, and dependency parsers. The abstractions are represented as a graph, and the terms were searched in the subgraph that satisfies certain global and local properties. The semantic graph generator is used to generate the graph where the questions and their answers are represented as a graph.

2.3 COLLABORATIVE LEARNING

Thanachai Wongvibulsin et al (2003) described a framework for collaborative question answering system for e-learning where the teaching assistant was actively participating to provide the answers for the students [12]. The student can send the message to the system and can receive the answer form the teaching assistant of the collaborative forum. The students from the different classes can also actively participate in the forum. It is required a communication agent and the protocol to communicate in the
system and the interaction protocol consists of the request, reply, check request, check reply and answer acceptant.

The different agents involving in this system are question agent, answer agent, checker agent, and the management agent. Each agent was allowed to ask the question as well as to provide the answers and the four protocols could check the correctness of the answers to the posted query.

Chun-Chia Wang et al (2006) developed a semantic based automatic question answering system with the support of collaborative learning where the students could participate in the system through online and the virtual teacher was giving response for the question posted by the student [25]. The teacher might have analyzed the problem of students in every aspect which helps the teacher to give the great response for the student’s query. It supported the group learning methodology which gives better results than learning alone. The student assistance agent was acting as the interface between the students and the QAS and Q&A Acquirer was storing the unanswered questions posted by the students and the semantic index module is used to maintain the semanticness of the system.

Fernando Bobillo et al (2011) described a question answering system using Fuzzy ontology representation using OWL 2 which incorporates reasoning with the support of fuzzy [54]. It used OWL 2 editors for fuzzy ontology representation and OWL 2 reasoner discard the fuzzy part of a fuzzy ontology. The fuzzy member function is applied to fuzzy modifiers which in the range between 0 to 1 and supports to modify the fuzzy member function. This research also focused on the fuzzy concrete domains which are having a set of predictors and applied in different members like trapezoidal membership function, the triangular, the left-shoulder function and the right-shoulder function.
Prof. Kohei Arai et al (2012) framed an effective question answering system for collaborative learning, which can act not just like a virtual teacher, but also the virtual discussion for student [53]. The student can attach their question when they want collaborative using collaborative learning capitalize on one another’s resources and skills. Students can ask their questions to the group when they want to collaborate with others, asking one another for information, evaluating one another’s ideas. The method also considers that students can communicate through Q&A interaction such a discussion forum to support information sharing.

2.4 SUMMARY

Based on the literature survey, existing works were categorized and analyzed by stating their advantages and disadvantages. Based on the findings, the problem definition and the methodology adopted for the question answering system were studied in a detailed manner. The objective of the research work has been identified and their previous methodology of the system has been analyzed. From the existing work of the question answering system, the requirement has been identified and it helps to define the objective of the proposed work. Based on those ideas the proposed system has been implemented with a new integrated approach.