CHAPTER 1

QUESTION ANSWERING SYSTEM

1.1 INTRODUCTION

Nowadays everyone around the world is trying to retrieve some information from the web. Search engine has done a fantastic job for retrieving information to the user from the web. The information which is retrieved from the web may be in structured document, unstructured document or links of other web pages. The retrieved information is listed using ranking algorithms. But, the user does not get the expected answer during searching. From the collection of document retrieved and listed in ranking order, the user is required to select the correct answer from the retrieved documents with specified links. It is a time consuming job and there is a possibility of choosing irrelevant information from the listed links. A question answering system comes into role for solving this problem.

A question answering system is quite similar to a search engine. Question answering system and search engine have user interface and are used to allow the user to find relevant data. Search engine relies on human to put queries and not questions. A query is a formatted string which contains keywords and search engine commands. The output of a search engine is a set of documents that match the users query. But, a QA system takes a natural language question as input and returns a specific answer as the output. Question answering system will not produce the list of links,
instead of that it produces the answer in few words or in single sentence form.

Question answering mechanism is a task of automatically answering a question posted in natural language. The main techniques used in QA system are Information Retrieval (IR) and natural language processing. To answer a given question, a QA may use either a pre-structured database or a collection of documents in natural language form. QA deals with different question types such as factoid based, list based, definition based, reason based and explanation based types. Factoid based question is based on the basic terms of wh-form question types which are represented in simple form. The successful functionality of question answering system depends on the type which is to be done from small collection of documents to large format of reports. QA system is used for the need of user to collect specific information and it does not has the time to read all available documents related to the search topic.

The organization of the chapter is as follows: Types of QA systems are explained in Section 1.2 and Section 1.3 presents the methods of QA system. Issues of QA and needed technologies used in QA are discussed in Section 1.4 and Section 1.5. History of QA and general architecture of QA are explained in Section 1.6 and Section 1.7. Future of QA is discussed in Section 1.8. Section 1.9 is used to explain the evaluation metrics of QA system.

1.2 Types of Question Answering System

1.2.1 Open Domain and Closed Domain QA

Nowadays two types of QA system are available and they are classified as open domain QA systems and closed domain QA systems.
Open domain QA system deals with a set of questions which can be asked anything and can lie on general information which is stored in a knowledge base. Knowledge base is a collection of terms which are represented with its meaning. It is like a database. But, there is a need of large collections of data for extracting the answer. The size of the database should be large for storing all the information in it and the type of stored data is not only in text form but also in multimedia form. The retrieval of correct answer from this large dataset takes more time. Effective searching methodology is needed for retrieving answer.

Closed domain QA deals with questions under a specific domain and it is an easy task. Because, NLP systems can exploit domain restricted form described by ontology. In this model, limited types of questions are accepted. The user should have knowledge on that closed domain to find the accuracy of the resultant answer. The size of database is small compared to open domain QA system.

1.2.2 Factors for Closed Domain QA

There are various factors available for using closed domain QA. The factors considered are size of data, domain context and resources. The first factor is the size of data which is used as a method in QA and it is derived from redundancy techniques. In this method answers were
can ask specific terminology rather than technical questions with specific answers. The experts also verify the resultant answer with the expected answer. This is only possible with in-depth knowledge in particular domain.

The third factor is a set of resources. It is an important difference between open domain and closed domain QA where the existence of resources used for closed domain is clearly available and easily accessible. Storage of data in knowledge base depends upon the type of resource.

1.2.3 Characteristics of Closed Domain

The nature of a closed domain affects the kinds of questions asked and the answers expected. Some domains are particularly suitable for the development of question answering system. If a QA system is not related to World Wide Web (WWW) and it is used within a particular domain, then there is a need of clearly defined knowledge source. The techniques used in a closed domain should not use extensive knowledge from outside domain.

A good method for answering questions in a closed domain requires forcing of any information available for the domain. The domain needs to be able to notify the user’s information requirements with
1.3 METHODS OF QUESTION ANSWERING SYSTEM

1.3.1 Shallow Parsing

Shallow parsing is an analysis of a sentence which identifies noun, verb, verb group and others. Shallow parsing is based on keyword based technique to locate answer from retrieved documents and the answers are filtered. Filtering mechanism is based on the presence of answer type. Ranking is done by using some features and similarity queries. Some system use templates to find exact answer with good data redundancy and there is a need for reformulation of question. Factoid based questions work under this method very well because factoid based questions mostly based upon name, thing, location, quantity and time. One of the main application areas of shallow parsing is question answering field. It provides the structural basis for the Natural Language (NL) questions. To analyze NL question, this shallow parsing technique will be used in a predefined manner.

Typical modules within shallow parser architecture include Part_Of_Speech (POS) tagging, chunking and relation finding. Tagging is a task of labeling each word in a sentence with an appropriate part of speech. POS tagging is used to find the correct class of the word such as noun, verb and adjectives. POS tag is used for questions with documents, tokenize text and disambiguate tokens based on context using WordNet tool. WordNet tool contains collections of English words in it. In chunking, a group of noun phrases, verb phrases and complete clauses are used. For a given set of words and their classes, this technique is used to decide the given words formed as groups. In relation finding, a given set of chunks are described as a subject, object and location. For a given set of chunks in a sentence, this technique is used to decide the relation in a
sentence. Shallow parsing is used in different applications such as speech-to-speech translation system, question answering system and text-mining application.

1.3.2 Deep Parsing

Deep parsing is not only based on query reformulation and keyword techniques but also suited for syntactic, semantic and contextual processing of answer extraction. The techniques are termed as Named Entity Recognition (NER), Word Sense Disambiguation (WSD), logic form and inferences. The conceptual graph formalism is a powerful semantic knowledge representation method which is used to model knowledge in documents and questions. In this type, comparison of two words is used by projection algorithms. Complicated queries like why or how questions and dialog queries comes under this method.

The main problem in deep parsing is to select the most possible syntactic analysis. Stochastic approaches can be used to order the analyses according to their probability or to generate the most probable parse. However, a complete syntactic analysis is not required for all NLP applications. A full parse often provides more information than the needed one. For example in IR deep parsing, it is enough to find simple Noun Phrase (NP) and Verb Phrase (VP).

1.4 MAJOR ISSUES

The major issues of QA are classified as question classes, question processing, context in question, data source, answer extraction and answer formulation.
In question classes, different types of questions are formed with different strategy to find the answer. The taxonomy is arranged for classifying the question. This classification mechanism should be taken automatically by the QA system. In question processing, a semantic model is needed to understand the user’s question and processing is done in it in correct form. The query analysis and reformulation of question is taken in this model. This question processing model enables the translation of complex question into series of simple question forms using interactive clarification from user.

In QA issue, context is used to clarify the question and resolve ambiguity for understanding of question. Logic or rule model is needed for understanding of questions. Rules should also be satisfied for all types of questions. In data sources, the QA system is in need to know the knowledge source for answer retrieval. Data sources take a role for storing knowledge about the data in knowledge base. More number of data is related to particular source that will produce more accurate answer. So, the answer will be automatically satisfied by the user.

Answer extraction method depends on the complexity of question by using searching method and focus of question. Searching method can be processed by either keyword based type or semantic based type. Keyword based type is also called as syntactic model. In keyword based model, exact keyword terms are matched with the terms available in the knowledge base. In semantic search based model, the meaning of terms is analyzed in knowledge base and the result is produced. But, semantic search model contains a complex design compared to keyword based type. In answer formulation, result retrieved from the knowledge source is in natural form. Simple extraction is sufficient for retrieval in
this answer formulation. The user can expect the result in text form, multimedia form or visualization form.

1.5 **NEEDED TECHNOLOGIES**

To implement QA in effective form, three major technologies are used namely machine learning, information retrieval and information extraction. Machine learning techniques for question answering are often based on small set of answer types like location, person and time. These are easily generated using named entity tagger. Answer types are usually arranged in a shallow parsing method at various levels. Information retrieval is a major task in QA system to inform the existence or non-existence of information in the knowledge source in accurate manner. Information Extraction is the advanced level of information retrieval process.

1.5.1 **Machine Learning**

The machine learning techniques are used in question classification module and answer extraction module. Witten & Frank (2005) have discussed the implementation of various machine learning algorithms using Waikato Environment for Knowledge Analysis (WEKA) tool for question classification. WEKA tool supports all type of classification techniques and answers can be verified with large data set. Data can be imported from a file in Attribute Relation File Format (ARFF). Data can be read from a Uniform Resource Locator (URL) or a Structured Query Language (SQL).

The field of machine learning is divided into three categories as supervised, unsupervised and reinforced learning. Supervised learning involves learning a function from a training set of pairs of inputs and its
corresponding outputs. The learning is said to be supervised because a supervisor is required to direct the learning process by supplying the desired outputs to the corresponding inputs. In an unsupervised learning, the goal is to learn the pattern and interesting structures directly from the input without knowing of corresponding outputs. In reinforcement learning, an agent learns how to act in environment by means of maximizing a reward function.

1.5.2 Information Retrieval

Information retrieval refers to Human Computer Interaction (HCI) that happens when the search is used in machine. Depending upon the programmed sophistication of machine, a user query is matched against a set of documents to find a subset of relevant information.

IR is relevant to question answering for two reasons. First, IR techniques have been extended to return not just relevant documents, but also relevant passages within documents. The size of these passages can be steadily reduced to produce answer to a question which is extracted from the passage. Thus, QA has passage retrieval in limit.

Second, the IR community has developed a methodology for evaluation which is the most well-known current example in the annual Text REtrieval Conferences (TRECs). These TRECs are run by the US national institute of standards and technology. A recent question answering system evaluation has been developed based on the TRECs community which in turn has stimulated much of the current interest in question answering.
1.5.3 Information Extraction

Information Extraction (IE) is a technique to analyze unrestricted text in order to extract information about pre specified types of event, entity or relationship. It is an advanced level of information retrieval process. It derives structured and factual information from unstructured text. Currently, it is applied in text mining field. IE refers to the ability of machine for automatically extracting structured information.

1.6 HISTORY OF QA

In 1990’s, TREC’s include a question answering track which is running till now. The best TREC’s system of 2004 achieved 77% correct answers for factoid based questions. Google and Microsoft companies had also started to integrate question answering technique with search engines. These tools are basically worked by shallow type of parsing methods. Starting in 1999, the Text REtrieval Conference-1 (TREC-1) had sponsored a question answering track which evaluated QA systems to answer factoid based questions by consulting the documents of the TREC’s corpus. A number of systems in this evaluation had successfully combined with information retrieval and natural language processing technique. The first QA was developed in 1960’s and was basically a natural language interface to specific domain. But in current QA systems, text documents are used to underline the knowledge and various natural language processing techniques are used to search for retrieval of answer.

Research in QA is in three different perspectives namely Artificial Intelligence (AI), natural language processing and ontology. In early stage of work in QA, artificial intelligence is mainly used to implement human computer interaction model. AI is used to respond to
question using knowledge which is encoded in database as information source. These types of systems can provide answers with the previous information encoded in database. For complex type of information, these systems are used as theorem proving and reasoning techniques.

Question answering has many applications which can be divided as applications based on source of answers in knowledge base, structured data model of knowledge base and semi or free structured data type.

1.6.1 Artificial Intelligence

The best known early question answering system is a Program for answering questions about BASEBALL games (BASEBALL) played in the American league over one season. Green et al (1961) has developed a QA system BASEBALL which is the first of a series of programs designed as natural language front ends to databases. BASEBALL is restricted to questions about baseball facts and most QA system is used for a long time. It is restricted to structured database model. In this closed domain QA, users are allowed to input queries in natural language then the interface is used to analyze the syntaxes and meanings by using linguistic knowledge.

The most well remembered other early work in this tradition is the Application to LUNAR geology (LUNAR) system. LUNAR is designed to enable a lunar geologist to conveniently access, compare and evaluate the chemical analysis data on lunar rock and soil composition that is accumulating as a result of the apollo moon mission.

Voorhees & Ellen (2001) have introduced TREC's question answering track which is motivated in the field of question answering. The
initial efforts in question answering is focused on fact-based and short-answered questions. It is aimed for comparing IR systems implementation by academic and commercial research groups. The TREC systems are run by pre-selected queries and retrieved text documents. The results are evaluated manually. This type of TREC systems is used by shallow NLP technique based on pattern matching algorithm.

1.6.2 Natural Language Processing

A natural language is a system for describing perceptions. Natural language is in natural form of interaction by human with a machine. NLP based QA provides the human computer dialogue and answers are in reliable ways. These NLP based QA produced answers which are achieved by mapping text into semantic real world representation in formal manner. The primary focus of NLP is the use of natural language which acts as an interface for databases in a narrow domain. IR is one of the sub modules of NLP.

IR systems use statistical methods and retrieve complete documents. But, the answer for the given question is represented in indirect manner. This is one of the limitations of IR. IR system does not attempt to understand the meaning of users query or documents in the collection. In contrast with classical information retrieval, complete documents are considered as relevant to the information request. In question answering, specific pieces of information are returned as an answer.

QA targeted different types of sources such as structured databases, unstructured text format and precompiled knowledge base. In IR perspective, question answering is focused for finding text that contains
the answer within large collections of documents. A specific kind of question answering is easy to evaluate and is focused on shallow based parsing methods which are independent of application domain. Structured knowledge based QA systems are adapted with complex queries in structured information environment. In conventional IR, the input is given in keyword format and the output is represented as a collection of document list. But in QA, a question with NL format is given as an input and the words, phrases which include the answer are formed as output. The input and output format differs in information retrieval and question answering systems as represented in Table 1.1.

Table 1.1 Information Retrieval and Question Answering

<table>
<thead>
<tr>
<th>Method</th>
<th>Information Retrieval</th>
<th>Question Answering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>Keywords</td>
<td>Natural language format of question</td>
</tr>
<tr>
<td>Output</td>
<td>Document List</td>
<td>Words, phrases with answer</td>
</tr>
</tbody>
</table>

1.6.3 Ontology

Ontology is defined as “a formal explicit specification of a shared conceptualization”. A conceptualization term in the definition refers to an abstract model of some occurrence in real world. An explicit word in ontology definition represents the type of concepts used and explicitly defines the concepts. Formal term in the definition refers a fact that the ontology should be in machine understandable format. Shared term represents the ontology which is not restricted to particular user but also for group. Ontology describes a shared understanding of a domain of interest. It is also represented as a formal and machine manipulation
model with domain of interest. For example, object oriented is taken as a closed domain ontology in which class, object, methods are represented as classes. Abstract class, friend class, virtual classes are the sub classes of ‘class’ concept. Like that relations ‘has_method’ is represented as relations between ‘abstract class’ and ‘abstract method’.

It is an essential technology to provide formal semantic understanding to a computer with real world semantic understanding to humans. Ontologies are developed in AI for sharing and reusing knowledge. Ontology is a data model that represents a set of concepts within a domain and the relationships between the concepts. It is used for reasoning about the objects within that domain.

With the following reasons only, the ontology development is taken in fast form. These are used to understand the basic structure of information among people, to reuse domain knowledge and to make the domain knowledge in explicit form. These reasons are achieved by defining classes in ontology, arranging classes in taxonomic hierarchy, defining slots and describe values for these slots with filling values for slots of instances. The knowledge base can be created by using individual instances of classes with filling specific slot value and restrictions.

Main components of ontology are properties, value restrictions, disjointness statements and logical relationships between objects. Ontologies play a major role in web because it provides a semantic interoperability to overcome differences in terminology. Ontologies are useful for organization and navigation of web sites. Ontology also used for web search because it is used for improving the accuracy of web search. It uses generalization and specialization information during web search.
Noy & McGuinness (2001) have listed the uses of ontology in question answering such as the ontology is used to share common understanding of the structured information among people. Ontology is used to enable and reuse the domain knowledge with the domain assumptions in explicit form. It is also used to separate domain knowledge from the operational knowledge and it is making different analysis of the domain knowledge.

Ontology concepts and resources are used as ontology file in Resource Description Framework/Web Ontology Language (RDF/OWL) format. QA approaches have been focused on retrieving answer from a raw text by using ontology to markup web resources and to improve retrieval with query expansion. QA systems can be classified based upon semantic search approach as four types. These types are the input or type of question, sources of information in structured or unstructured form, domain specific or domain independent and adaptability of non-trivial systems. Ontology based QA system takes queries in NL form with ontology as input and ontology returns answer from Knowledge Base (KB). KB subscribes the answer to the ontology. So, there is no requirement for the user to learn the vocabulary or structure of ontology to be queried for getting the answer.

1.7 QA SYSTEM ARCHITECTURE

Question answering system mainly focused in three phases such as question classification, query processing and answer extraction.
Figure 1.1 General Architecture of Question Answering System

The Figure 1.1 shows the general architecture of QA system with these three modules. Question classification module is used to classify the type of question which is given by the user and its possible type of output format. In this module, the type of question and answer format are finalized. Logic rules are implemented using a machine learning approach for removing stop words in question and these rules are used for getting correct answer. Stop words are commonly used words which do not contain important significance to be used in search queries. A search engine has been programmed to ignore these stop words.

Query processing module is used to convert the given type of question into query form for finding out the answer from ontology. Reformulation of query is also possible in this module. Finally, answer extraction module is used to retrieve the answer from ontology. The
answer is represented as concept or relation or instances in ontology. Knowledge base is a database which holds collection of terms in either structured or unstructured format.

1.7.1 Question Classification

Question classification module is the first module in QA system which is used to determine the type of question and the type of answer. After question analysis module, system uses NLP techniques to reduce the amount of text. The accuracy of question classification is very important to the overall performance of the question answering system. In this module, common words like ‘what’, ‘where’ question terms and ’is’, ‘was’ like verb terms should be noted for classification. Stop words are irrelevant for searching and these words are needed to be dropped at indexing time. Common stop words used in question are ‘the’ and ‘a’. It is possible to manually write some rules for question classification. But, it is a tedious work to implement in real time environment. So, a machine-learning approach can be used to automatically construct a classification method.

1.7.2 Query Processing

The second module is the query processing module which is used to support search in ontology. Search is used to match the given question with the specific ontology structure for retrieving the answer from ontology. To process this module successfully, parsing takes a major task for conversion of given keywords into a token form for searching process. Matching between query and specific ontology subgraph is based on semantic similarity matching mechanism. The question pattern for each question Q is identified by using statistical and linguistic information.
Here, a Question pattern $Q_p$ is defined as a question word and one or two keywords that are related to the question word. $Q_p$ represents the question intention and it is indicated for named entities.

The conversion of question to query form is processed after reformulation of given question. The reformulation of question is processed with the help of user and with WordNet only when the expected answer is not produced by ontology. Sometime, query expansion can also be processed for effective retrieval of given question. In ontology based QA, Simple Protocol for RDF Query Language (SPARQL) is used as query processing language. This SPARQL is used for supporting the unification, bisection, differentiation rules. This language can be used like SQL.

1.7.3 Answer Retrieval

Answer extraction module takes the given answer from ontology with the satisfaction of the user. In a factoid question framework, the answer is in general form which is limited to a portion of a text. From that portion of text, only the answer has been found. This may be sufficient for factoid questions when the answer does not need any elaboration as prior. But, this is not appropriate for the types of questions presented in other format of questions like list based, definition based, reason based and explanation based type.

Answer generation is a complex and critical task which has seldom been addressed in the question answering community. As in other areas of language processing, most of the language generation aspects have been neglected. The result is that most applications may have a rich question analysis facet with a poor output and possibility of leading to
misunderstandings. So, there is a need of effective answer retrieval technique for correct retrieval of answer.

1.8 FUTURE OF QA

QA systems are extended in recent years to explore new dimensions in scientific world. Research is in progress to automatically answer temporal and geospatial questions, biographical questions, multilingual questions and multimedia questions. The following concepts are needed for future research of QA which are answer reuse, interactivity, knowledge representation and reasoning. Automatic question answering is performed using a single resource to answer user question. It is in need to have rich knowledge base that should be updated constantly.

1.8.1 Semantic Web

“Semantic web is a machine readable information and automated services that extend far beyond the current capabilities” told by Lee et al (2001). Semantic web is implemented with the help of ontological support for representing database. Semantic web not only based on text based manipulation but it uses machine process based metadata format. Semantic web uses intelligent techniques to take advantage of machine processing representation. This type of web needs a standard format for storing arbitrary data. It is an abstract representation for various types of data. To represent semantic web, RDF language is used. RDF is a mechanism to tell something about the data. RDF data model is a simple one which consists of resources and triples. The combination of class, property and object is called as triples. RDF Schema (RDFS) is a simple type of RDF system which defines properties and classes that are applied to particular domain.
Lopez et al (2011) have discussed the comparison of question answering system with semantic web community in a detailed manner. In this paper, the main goal of natural language based QA system is used to answer queries by locating and combining information which are distributed massively as heterogeneous resources. These resources can be represented by using semantic web.

In Figure 1.2, the semantic web architecture is noted as layered format. Semantic web consists of web enabled languages that allow machine processed semantics of data. It also contains a set of rules for designing the query language for getting the relevant answer from the web. The different types of query languages used for answer retrieval are discussed by Ding et al (2002). These formats are discussed as follows.

Figure 1.2 Semantic Web Architecture
1.8.1.1 XML

The eXtensible Markup Language (XML) is a meta language that is used to define application with specific markup tags. XML has a mechanism for standardized representation of other languages. It provides a data format for structured documents. A Document Type Definition (DTD) or XML schema might be used to specify the vocabulary and define multiple tags. Sometimes, the meaning of piece of XML tags may seem to human users. But, all are formally specified in general form. It contains named entities with values and sub entities. Every XML document form is in ordered tree format with labeled manner. It is a major strength and also sometime it is a weakness. It is possible to represent all type of data structures in an unambiguous syntax form.

1.8.1.2 RDF/RDF schema

XML is a standard mechanism for structured data and not for unstructured form of data. So, RDF comes into a role for a mechanism of defining unstructured form of data. RDF is a model which tells something about the meaning of data. It is not a language, but a model used to represent data about “meaning of data”. This type of meaning of data is called metadata. Meaning represents the resource used in RDF vocabulary. RDF is a data model originally used for describing metadata for web pages and it contains structured information in it. It is a universal and machine readable data exchange format which contains main syntax with XML for serialization.

Klyne & Carroll (2003) have discussed this RDF model with syntax specification in their work. Resources are the identifiers on the web from specific parts of documents and identical parts of document that is
used for persons or companies. RDF is a directed, labeled graph data format for representing information in the web.

RDF data model is simple which consists of statements about resources, encoded as object-attribute-value triples. Resources are the web pages represented in RDF format. The objects are resources or strings. Attributes are the properties of resources. RDF Schema is a simple type of system for RDF. It is a vocabulary description language. RDFS describes properties and classes of RDF resources. RDFS also provides semantic for generalization hierarchies of properties and classes. It provides a mechanism to define domain specific properties and classes of resources to which the properties can be applied. RDF and RDF Schema provide a simple knowledge representation for web resources. It is a full-fledged knowledge representation language.

RDF Schema consists of class definitions and sub-class of statements which are used to allow the definition of class hierarchy. RDF Schema has property definitions and sub-property of statements that are used to build property hierarchies. It also contains domain and range of statements used for restricting the possible combinations of properties and classes. The different types of statements are used to declare a resource and instance of a specific class.

1.8.1.3 OWL

Web Ontology Language (OWL) is a standard ontology based language for semantic web. It includes conjunction, disjunction, existentially and universally quantified variables. Reasoners can use of this to carry out logical inferences and derive knowledge. A reasoner is a piece of software which is able to infer logical consequences from a set of
asserted facts. It is in three forms which are described as OWL FULL, OWL DL and OWL Lite. OWL FULL is an upward compatibility with RDF both at syntax and semantic level. It is possible to change the RDF primitives. OWL DL is less time-complex. It allows efficient reasoning and inferencing. But, it loses backward compatibility with OWL FULL. OWL Lite is an even more restricted subset of OWL FULL for an expressive ontology language with decidable inference.

The different version of this ontology language is given with a choice to select the best one depending upon the system requirements. Depending upon the OWL type model, reasoners are used for checking the inference mechanism.

1.8.1.4 SPARQL

SPARQL can be used to express queries across diverse data sources in RDF form. SPARQL contains capabilities for querying required and optional graph patterns along with their conjunctions and disjunctions. It is like SQL query. It also supports extensible value testing and constraining queries by source RDF graph. The results of SPARQL queries can be a result set or RDF graph.

There are four types of query result formation available in SPARQL such as SELECT, CONSTRUCT, DESCRIBE and ASK. SELECT is like a general SQL form of select query type which can hold condition in it. FILTER command is used for restrictions in query. These restrictions are represented with regular expression type with Question symbol (?). The restrictions contain a constant value which may be an integer or character constant type. The restriction is also an optional one in SPARQL query.
CONSTRUCT form specifies a graph that to be returned with variables which are substituted from the query pattern. DESCRIBE form will return information about matched resources in RDF graph type. The exact form of this information is not standardized but usually a blank node closure is expected. ASK form is intended for asking yes/no questions about matching. If no information about matched variables is returned, the result will indicate whether the matching exists or not. The sequence of result can be modified using the following keywords like ORDERBY, DISTINCT, OFFSET and LIMIT.

1.8.1.5 Proof, trust and logic

Proof layer is used for generation, exchange and validating data. Trust layer supports security among web with digital signature. Logic layer enhances ontology languages with the application of specific declarative knowledge.

1.8.1.6 URI/IRI

Uniform Resource Identifier (URI) is a string of computers used for identifying name or a web resource. This type of identification enables the interaction with web resource in WWW. URI can be classified as locators and names. Locator provides a method for finding identity and names define items identity. Internationalized Resource Identifier (IRI) is a general form of URI which is used in internet. URI is limited to subset of American Standard Code for Information Interchange (ASCII) character set where IRI may contain characters from universal character set. It is not too difficult for anyone to replicate arbitrary UNIversal CODE (UNICODE) on keyboard.
1.8.2 WordNet

A shallow semantics provide a system which contains a relationship of words with synonyms and connection among these synonyms. But, as the name indicates it could not perform a deeper inference like processing a causal relation or other logical deducing. An example of semantic network is WordNet, an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets and each one is used for representing the underlying lexical concept. Different relations link the synonym sets.

WordNet is a lexical database in which words are grouped into sets of synonyms called synsets. Synsets are connected together by means of semantic relations such as hypernym and antonym. Hypernym is a word that is more generic than a given word. It describes things more specifically. Proper nouns are good example of hypernyms. Antonym is a word that has the opposite meaning. It refers to a word that is completely different from another. WordNet stores a set of synsets, hypernyms and antonyms for a set of words in it. WordNet is used to provide more intuitively usable combination of dictionary and thesaurus. It supports automatic text analysis and artificial intelligence.

1.8.3 Tools for Ontology Construction

Incorporating the methodologies and languages, a system is needed for building knowledge acquisition. Protege tool is used for satisfying this condition. Its main feature is the inclusion of extensible knowledge model to enable users and to redefine the representational primitives. It is a powerful plug-in architecture to enable integration with
other applications. It generates the output in many ontology languages and it is easy for users to change it to an editor of a specific language.

Protege is a free open source ontology editor which allows generation, visualization and manipulation of ontologies. Ontology is created and it can be accessed from java programs through the protege OWL Application Programming Interface (API). It represents classes, properties and individuals. Things are represented as class in ontology. Properties have object type and data type.

1.8.4 Semantic Search

Semantic search is used to improve the accuracy of search by understanding the intent of the user and the meaning of the terms in sentence. There are two types of searches available such as navigational search and research search. Research search is also called semantic search.

In navigational search, the user uses the search engine as a navigation tool to navigate to a particular intended document. Semantic search is not applicable to navigational search. In navigational search, the user provides a search engine with a phrase which is intended to denote an object about the user who tries to gather or research information. Rather than Google’s pagerank algorithm, semantic search uses semantics to produce highly relevant searching results. This semantic search technique can be used to retrieve the knowledge from the data source like ontology. Ontology is a technology used to enable the domain knowledge at a high level and improve the query time.
1.9 EVALUATION METRICS

In order to analyze the overall performance of question answering system, many types of evaluation metrics are available. Mostly in QA system accuracy, precision, recall, F-measure and MRR metrics are used for measuring the performance of the system. These metrics are used at various modules in question answering system such as query classification and reformulation module, query processing module and answer extraction module. In this thesis for answer extraction module, confidence ratio is also used as an important factor for finding the resultant answer is correct or not. These popular measurements in question answering system are listed below.

Precision is the fraction of retrieved instances that are relevant while recall is the fraction of relevant instances that are retrieved. Precision can be seen as a measure of exactness or quality, whereas recall is a measure of completeness or quantity. In simple terms, high recall means that an algorithm returned most of the relevant results. High precision means that an algorithm returned more relevant results than irrelevant. Precision and Recall are set of measures which are used to evaluate the quality of an unordered set of retrieved answers. For evaluation, precision can be plotted against recall.

1.9.1 Accuracy

Accuracy is measured by a number of questions answered correctly by the total number of questions asked in a test and it is represented in Equation (1.1).
In questions answering system, accuracy is measured by a main evaluation metric.

### 1.9.2 Precision, Recall and F-Measure

Precision is defined by a number of correctly answered questions by number of returned answers and represented in Equation (1.2).

\[
Precision = \left( \frac{\text{number of correctly answered questions}}{\text{number of returned answers}} \right)
\]

(1.2)

Recall is defined by number of correctly answered questions by number of possible correct answers which is described in Equation (1.3).

\[
Recall = \left( \frac{\text{number of correctly answered questions}}{\text{number of possible correct answers}} \right)
\]

(1.3)

For classification tasks, the terms True positive (tp), True negative (tn), False positive (fp) and False negative (fn) compare the results. True positive means correct result from the answering system. False positive is calculated based on unexpected result for the given question. True negative means correct absence or result of answer and false negative means missing result for the given question. So according to
this classification scheme, precision and recall can also be measured in Equations (1.4) and (1.5) respectively.

\[
\text{precision} = \left( \frac{tp}{tp + fp} \right)
\]

(1.4)

\[
\text{recall} = \left( \frac{tp}{tp + fn} \right)
\]

(1.5)

F-measure combines precision and recall measures as a single unit. It is a harmonic mean of precision and recall which is represented in Equation (1.6).

\[
F - \text{measure} = \left( \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \right)
\]

(1.6)

1.9.3 Mean Reciprocal Rank

The mean reciprocal rank unit is the first used in TREC's question answering system. It is used to evaluate the system which returns a list of possible answers for a given question in ranking order. It is defined as a multiplicative inverse of a rank of first correct answer. It returns zero when no answer is returned. It is an average of a reciprocal rank of every question in a test set.
In Equation (1.7) $rank_i$ denotes rank of every question in $i^{th}$ test set.

1.10 ORGANIZATION OF THE THESIS

The organization of the chapter 1 is as follows: Types and methods of QA systems are explained in Section 1.2 and Section 1.3. Issues of QA and needed technologies used in QA are discussed in Section 1.4 and Section 1.5. History of QA is discussed in Section 1.6 and general architecture of QA is explained in Section 1.7. Section 1.8 discussed the future of QA. Section 1.9 is used to explain the evaluation metrics of QA system.

The organization of the chapter 2 is as follows: Section 2.2 is explained by natural language based QA. Ontology based QA and semantic search in ontological QA are discussed in Section 2.3 and Section 2.4. General principles of AQUALOG QA system is discussed in Section 2.5. The problem statement and objective of the proposed work are described in Section 2.6 and Section 2.7.

The organization of the chapter 3 is as follows: QAAL system architecture is explained in Section 3.2. Analyses of question pattern and query formulation are described in Section 3.3 and Section 3.4. Question to query algorithm in QAAL system is explained in Section 3.5. Section 3.6 discusses the implementation of QAAL system with Q2Q algorithm. Sample query usage in QAAL system is explained in Section 3.7. In
Section 3.8, the performance evaluation and discussion is based on precision, recall, accuracy and recognition ratio values. Finally concluding remarks as summary is made in Section 3.9.

The organization of the chapter 4 is as follows: The survey of semantic search based QA system is discussed in Section 4.2. Pattern matching techniques and semantic search based matching terms are explained in Section 4.3 and Section 4.4. Section 4.5 and Section 4.6 are discussed about the reasons for graph representation and conceptual graph. RDF graph and semantic search matching with RDF graph are described in Section 4.7 and Section 4.8. QAAL graph traversal algorithm and usage of semantic search based search with QGT algorithm is QAAL system are described in Section 4.9 and Section 4.10. In Section 4.11, the performance evaluation and discussion is based on precision, recall, F-Measure and MRR values. Finally concluding remarks as summary is made in Section 4.12.

The organization of the chapter 5 is as follows: Semantic indexing and semantic indexing in QAAL system are explained in Section 5.2 and Section 5.3. In Section 5.4 and Section 5.5, RDF graph matching with synonym factor and RDF graph matching with inferences using semantic indexing are discussed. In Section 5.6, the performance evaluation and discussion is based on synonym factor, semantic indexing, semantic search with semantic indexing and confidence ratio. Finally concluding remarks as summary is made in Section 5.7.

The organization of the chapter 6 is as follows: In Section 6.1 and Section 6.2, the conclusion and scope of future extension have been discussed respectively.