CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

All users are struggling to get a correct answer from online resource. There is a need of a system which allows the user to ask a question and receive answer quickly. The system should validate the answer automatically. Current answering systems do not deliver answers to the user exactly. A system must analyze the question with the preference of user and it must find the answer in resources for an answer of a given question. It should present the answer to the user in an appropriate form. Generally, question answering system performs a match to the query with existing queries from knowledge base. It generates a set of keywords or a set of queries in knowledge base and determines the result that matches to that question.

Implementation of QA may be differed as natural language form, ontology based QA and semantic web based QA. Each system has some merits and demerits in it. This chapter summarizes all the above types of QA with their approaches and merits, demerits of each types of QA is also discussed in this chapter. The organization of the chapter 2 is as follows: Section 2.2 is explained 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.

2.2 NATURAL LANGUAGE BASED QA

Natural language based QA system contains some tasks which are mentioned as automatic summarization, discourse analysis, machine translation, morphological segmentation, NER, Optical Character Recognition (OCR), parsing, speech recognition, stemming, text simplification, query expansion and speech processing.

Hirschman & Gaizauskas (2001) have analyzed a natural language question answering view in the following way. IR is mainly used in NL based QA to retrieve only the documents and not the answers to the user. The user extracts answer from the documents. Most natural language front end query is used as a subset of controlled language which is limited to vocabulary and syntax. It is implemented in explicit form to the database. Most NL analysis follows two steps. The first step is to identify semantic type of entity in the question like date, person and company name. The second step is to determine additional constraints in the answer entity like identifying keywords in question which is used to match the answer with a sentence and was used to identify relations between answer entity and events noted in question.

Some of the early Natural Language Interface to Data Bases (NLIDB) use pattern matching techniques. The main drawback of early NLIDB systems is that they are designed for particular database. It cannot be easily modified with different databases and difficult to port to different application domains. It has been taken long time because of the need of handwritten mapping rules for particular domain. The next generation
NLIDB system is an intermediate representation language which is expressed the meaning of users question in terms of high level concepts and independent of database structure. A formal semantic approach is used to follow this concept with high degree of portability.

Popescu et al (2003) have discussed a PRECISE system to map question to query (PRECISE) which is used to map questions to corresponding SQL query by identifying classes of questions. Questions are translated into sets of attribute-value pairs and a relation token is either for attribute or a value token. A lexicon is also used for synonym. Questions with unknown words are not semantically tractable and are not handled in PRECISE system. This is the major disadvantage of this system.

Kim & Kim (2008) have proposed QA system which give a set of candidate query question for users information need and these candidate questions are automatically generated from sentences. In this model, sentence split, named entity recognition, question generation, question filtering and answer indexing are used. For query generation, NER technology is used. Two problems are analyzed in this paper. First one is the generation of question from a given sentence and the second problem is to isolate meaningful sentence from generated questions. Two types of questions are used namely sentence question and entity question.
Samarakoon et al (2011) have described a closed domain QA system for automation of customer helpdesk. The syntactic, lexical and morphological variations are used for synonym over knowledgebase. Query understanding method is used for ranking. A genetic algorithm method is used for updating ranking parameters for changes in user queries. This genetic algorithm is proposed for updating the optimum values of weighting factors for adaptation of user queries.

Depending on the evaluation, three major types of evaluation is processed in NLP based QA system. Those are intrinsic and extrinsic evaluation, black box and glass box evaluation and automatic-manual evaluation. Intrinsic evaluation is characterized with its performance with respect to evaluators. Extrinsic evaluation is based on system with precise function for human user.

The main drawback of this NL based QA system is the lagging of semantic mapping of words in question for retrieval of correct answer. For implementing semantic representation of terms in question, ontology based QA system is suggested. The meanings of terms in questions are considered in ontology based QA system by using class, property and relation. The metadata construction is also possible in ontology. Inference and indexing methodology are used for successful retrieval of correct answer in QA system.
In ontology based QA system, knowledge based data with the answers has structured organization defined by an ontology. Users could raise questions in natural language and the system will return accurate answers to users directly after analyzing the question, retrieving information and extracting the answer from it. Ontology knowledge base provides a convenient way to obtain knowledge for users but, the natural language need to be mapped to the query in ontology.

Main challenges are the intention of user question in question analysis phase and another one is the mapping of user intention to the ontological query statement. The intention of user question can be solved through natural language processing technology and the mapping of user intention to the ontological query is crucial for ontology retrieval to get an answer. There are some relevant research work is going on related to QA system based on ontology model.

2.3.1 Survey of Ontology Based QA

Vargas & Motta (2003) have discussed A QUestion Answering system (AQUA) for heterogeneous sources for closed domain environment. It tries to answer a question using knowledge base and NLP technique is used. In this model, annotations are written in RDF notation that is a basic framework for expressing meta data. AQUA is used to translate english questions into logical queries which are used to generate proofs. Inference engine is used in this system based on resolution algorithm. The main contribution of AQUA is the use of ontology in various parts of QA system. In this model, ontology is used for the refinement of initial query, reasoning process and in similarity algorithm. Similarity algorithm is a key mode in AQUA which is used to find similarities between relations in translated query and relations in
ontological structure. AQUA does not require user intervention for building the graph of the query and the graph obtains from ontology. But, this system does not handle a set of follow up questions related to initial question.

Basili et al (2004) have discussed a Model-based Overlapping Seed ExpanSion (MOSES) QA system in which users post a question in natural language to knowledge base of facts extracted from a website and organized in topic map repositories. This approach is used by ontology based method to search, create, maintain and adapt semantically structured content in domain related type.

A question answering system is described by (McGuinness 2004) in semantic web model in detailed manner. In this article, ontology based enhanced processing and augmentation method is used for retrieving answer. Ontology enhanced processing and augmentation method is used to enhance answers with objects of a query. Another method is pruning for representing answer from semantic web. With this approach, the pattern matching is used to filter patterns with classes. Class is used for holding many properties with variables. Binding these variables is returned for pruning specification that is stored in ontology. Using this method, even simple QA system with limited knowledge can be improved its ability to find appropriate answers.

Hu et al (2005) proposes a web based QA system for closed domain. A probabilistic technique and a semantic analysis based on ontology are used. Okapi similarity measurement is suggested for retrieval of answer from ontology. Okapi is a method used in IR field for measuring similarity of words in sentence with frequency of query terms and average length of sentence. Semantic similarity between words is
computed and semantic annotation is also suggested in this model. NER is also implemented for measuring MRR values of answers.

Zhang et al (2006) have introduced a domain specific ontology for chinese language using ontology. Natural language parser is used to transfer sentences into semantic frame and reasoned. Query matcher is used to find the answer from the domain specific QA. Reasoner and query matcher are used to get main concept of question from concept catcher and query matcher uses a rule based classification mechanism for generating patterns.

Kaufmann et al (2006) proposes a Natural Language Interface to QUERy ontologies based on Clarification dialogS (QUERIX) system which combines stanford parser with WordNet to obtain RDF triples from natural language user queries. This approach is allowed queries in natural language. It consists of seven parts namely user interface, ontology manager, query analyzer, matching center, query generator, dialog component and ontology access layer. This approach is simple and does not have any complex queries in semantic manner and it does not resolve natural language ambiguities. It allows user to enter a question in NL query form and the output of the system is SPARQL query. Stanford parser is used in this system to provide a query in syntax tree form. From this syntax tree, the query analyzer is used to extract main word categories as noun, verb, preposition, wh word and conjunction. Matching is done as skeleton of keywords with synonym in triples form.

Cimiano et al (2008) have introduced an approach pORrtable nAatural language interface to KnowLEdge Base (ORAKEL) for computing intentional answers from a set of extensional answers which are returned as a result for the user query. Intentional answers are the
answers in terms of properties shared and enhances in underlying knowledgebase. In this approach, intentional answer is represented by a clause and computed based on Inductive Logic Programming (ILP) techniques. This approach is evaluated based on usefulness and time performance. The input is a natural language interface which is used to translate wh-word questions into logical queries. Evaluation is processed in logical queries based on given knowledge base and ontology with inference engine. Intentional answer describes the answer in hypothesis form is used to cover all positive answers and it is to be found in search space of program. Generalization algorithm is also implemented in this system for generating a hypothesis for covering extensional answers. This approach is a clause of literals that use non-recursive definite program clause. This system uses a reduction algorithm to evaluate the answer for the given query. Using this algorithm, irrelevant literal from the query is removed and it also increases the number of negative answers.

Wang et al (2007) have discussed a system called PortAble Natural language inTerface to Ontologies (PANTO) which is accepted a generic natural language queries and output as SPARQL queries. It is based on triple based data model to interpret the parse tree output by an off-the-shelf parser. Complex modifications in natural language queries such as negations, superlative and comparative are investigated in this system. It utilizes an off-the-shelf statistical parser called stanford parser. PANTO is designed as a portable ontology model for any specific domain in which WordNet and string metric algorithms are used for mapping the entities in NL queries. Nominal phrases in parse trees are extracted to form an intermediate represented called query triples. PANTO maps query triples to onto triples which are interpreted as SPARQL and represented as entities in ontology. The translations of advanced semantic features such
as negation, comparative and superlative modifications are supported in this system.

Lopez et al (2007) have discussed an Aqualog system. Semantic markup is achieved by using General Architecture for Text Engineering (GATE) NLP platform, string metric algorithms and ontology based similarity services for relations and classes. It has a plug-in mechanism which is used to allow different Knowledge Representation (KR) languages. Semantic search is processed in this system by using triple-based mechanism. There are two main reasons for using triple-based data model.

First, it is possible to represent most queries as triples. Second, RDF based KR model is used for database. For implementing triple based data model, AQUALOG uses linguistic component and relation similarity service. Linguistic component phase is completely domain independent and it is based on GATE analysis of English language. In this phase, the input question is converted into query form. Using GATE grammar query is termed as triple set. The output of this phase is the query-triple set form of Natural Language Query (NL-query). Relation Similarity Service (RSS) is the backbone of QA system which invokes NL-query and transformed as term-relation form for extracting exact result from database. For this, string metric similarity algorithm is implemented.

Battista et al (2007) have discussed Semantic web information Management with Automated Reasoning Tool (SMART) which is used to distribute biological knowledge in heterogeneous method. In this method, triples and DL reasoners are used for automated reasoning. It has features like semantic query composition, validation of query using DL reasoners, mapping DL queries to SPARQL for query conversion and retrieve
inferences from RDF triple. SMART is handled any type of ontology in OWL format for semantic query answering.

Ou et al (2008) have presented an automatic question pattern generation method for QA system based on ontology with the use of textual entailment. In this method, predictive questions are identified in a particular domain and these predictive questions are generated based on domain ontology. Query templates are used to extract answers to the predictive questions from knowledge base with the use of textual entailment engine. But, it has some drawbacks in it. Generated question patterns cannot work for all kinds of user questions and some generated question patterns does not make sense. There is also a limitation with the use of entailment method.

Ferrandez et al (2009) have discussed ontology-based Question Answering system applied to CInema Domain (QACID) which is used to retrieve information from formal ontology using input queries formulated in natural language format. That collection of queries is grouped as clusters where each cluster contained an alternative form for the same query. It contains multilingual capability, inter-domain portability and changes in user information requirements. In this model, data are stored in RDF type. This RDF is used as structured database used to extract answer by using SPARQL queries. Entities in ontology like classes, properties and instances are extracted from data directly and stored for fast access. Users query patterns are formed as SPARQL queries and these queries are permitted to access the information from RDF database.

Entailment engine is the main core system of this model which is used a technique to infer semantic deductions in user query and query pattern. Queries are processed using morphological analysis and Named
Entity (NE) tagging. Fuzzy matching technique is applied to match query with the ontology lexicon. Answer retrieval is in two stages. First, query pattern input is processed by entailment engine which is used to determine the semantic implications in query and patterns. Second, the SPARQL query is permitted to the expected answer from RDF database. SPARQL generator replaces the ontology concepts with data instances appearing in original query.

Damljanovic et al (2010) have discussed a natural language interfaces to ontologies named Feedback Refinement and Extended vocabulary Aggregation (FREYA) which has syntactic parsing and ontologies which contain knowledge in encoded form. It has high precision and recall. In this system, there are no specific rules that are framed for understanding the question. Instead of that, knowledge is encoded in ontology for accessing and understanding the question. This system not only resolves ambiguity but also maps the terms of questions with ontology concepts. It is also suitable for a set of grammatically incorrect questions. Questions are formed as semantic meaning and the answers are in concise manner. It has three steps as identification and verification of ontology concepts, generating SPARQL and identification of the answer type and presenting the results to the user.

Athenikos & Han (2010) have reviewed the current state of biomedical QA system based on semantic knowledge based QA approaches. According to this analysis, QA process is classified as semantic based, inference based and logic based methods. In knowledge based QA system, the knowledge of semantic information is extracted from textual sources and may fed in QA system for processing phrases to improve the performance of QA. Most semantic based QA approaches is taken as lexico semantic information encoded in WordNet as a resource
for the particular domain. In inference based QA, semantic relations are taken as inference between concepts and some inference rules are framed for improving the performance of QA. In logic based QA, explicit logic is formed and theorem proving techniques are used for transformation.

Moussa & Kader (2011) have discussed a Question Answering System for Yago Ontology (QASYO) which is used as sentence level QA system with integrated NLP, ontology and IR technologies. It is used as a triple based model and the binary relationships are used. Relation between the questions word are detected using parse tree split up with a candidate word combination. Questions are collected on the basis of true knowledge with linguistic restrictions. It is considered as adjectives, adverbs, nouns and verbs in question. This is a base line of search engine for specific query formation. Some words are determined experimentally that are acting as functional words as either logically or structurally.

Suchanek et al (2007) have discussed a semantic search based ontological system Yet Another Great Ontology (YAGO) which is used to integrate conceptual hierarchy of WordNet dictionary with Wikipedia. It is the first approach to combine WordNet and Wikipedia with the accuracy of 95%. YAGO approach utilizes the fact that Wikipedia has category pages. All objects like cities, people and URL are represented as entities in YAGO model. Two entities can stand as relation. YAGO has dozens of relations like taxonomic relation and non-taxonomic relation.

The aim of this section is to compare different types of ontological systems with ontology and semantic web based environment.
2.3.2 Limitations of Ontology Based QA

The Table 2.1 shows types of systems and its performance metrics in detail. The main limitation of comparison of all types of ontology based QA system is the user interaction. The user interaction is always needed for answer retrieval. It is a time consumption job. Query formation is not clearly defined in these set of QA systems. Most of the QA systems are not suitable for multi ontology open scenario. It is also required a domain specific lexicon. A few QA system is domain dependent, costly and can be applied to only to limited domain coverage. A closed domain QA system does not try to resolve NL ambiguities. But, it is also asked the user for clarification as pop up dialog window to disambiguate and the query to ontology is temporarily stored in a memory.

Table 2.1 Comparison of Types of Question Answering System with its Performance Analysis

<table>
<thead>
<tr>
<th>Types of QA System</th>
<th>Query Entry</th>
<th>Query Expansion/ Reformulation</th>
<th>Output Format</th>
<th>Success Rate</th>
<th>Domain Independent</th>
<th>Performance Metric Used in System</th>
</tr>
</thead>
<tbody>
<tr>
<td>PANTO</td>
<td>NL question</td>
<td>Synonym expansion</td>
<td>Answers</td>
<td>88.05</td>
<td>Yes</td>
<td>Run against ontology F-measure</td>
</tr>
<tr>
<td>FREYA</td>
<td>NL question</td>
<td>No</td>
<td>Graph Form</td>
<td>92.4</td>
<td>Yes</td>
<td>Precision, Recall and MRR</td>
</tr>
<tr>
<td>QUERIX</td>
<td>NL question</td>
<td>Synonym</td>
<td>Answers</td>
<td>86.08</td>
<td>Yes</td>
<td>Precision and Recall</td>
</tr>
<tr>
<td>ORAKEL</td>
<td>F-logic and SPARQL</td>
<td>No</td>
<td>Answers</td>
<td>93</td>
<td>No</td>
<td>Time and number of iterations for computation</td>
</tr>
<tr>
<td>QACID</td>
<td>NL query</td>
<td>No</td>
<td>Answers</td>
<td>80</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>AQUA LOG</td>
<td>NL question</td>
<td>Synonym expansion</td>
<td>Answers</td>
<td>63.5</td>
<td>Yes</td>
<td>User based F-measure</td>
</tr>
<tr>
<td>SMART</td>
<td>SPARQL</td>
<td>No</td>
<td>Answers and Query</td>
<td>80.3</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
Based on the performance analysis of different types of ontology based QA system, AQUALOG QA system is taken for this thesis work. AQUALOG is a portable question answering system which has less configuration time and different strategies are combined together in a novel way. It uses GATE NLP platform with string metric algorithms. The performance analysis is purely based on F-Measure ratio which combines both precision and recall values. FREYA system is also based on precision and recall values where the output is in graph form. The user does not understand the graphical taxonomy format of answer. PANTO and QUERIX systems are based on precision and recall values. But, stanford parser is implemented for query model. The query is displayed in syntax tree form on those systems. Among the study of all types of QA systems, AQUALOG QA system is taken for further study in this thesis work.

2.4 SEMANTIC SEARCH IN ONTOLOGICAL QA

In a keyword based question answering system, there is a possibility of high recall and low precision. The results which come from this type of system are highly sensitive to vocabulary. Human involvement is necessary for interpreting and combining results. There is a difficulty for distinguishing the meaning of two sentences in which keywords are same and these sentences is differred in meaning. This is due to lack of semantics in it. So, semantic search come into a role.

Semantic search is an application of semantic web in information retrieval for improving the performance of retrieval. Querying is applied in semantic data for implementing the semantic search. Several querying languages are applied for semantic querying. This semantic search is applied with semantic querying interfaces in four ways which is discussed by Uren et al (2007). These are keyword-based, form-based,
view-based and natural language based systems. Among these types, keyword-based query interface is more user-friendly. To combine keyword-based interfaces with semantic technologies is one of the challenging areas in semantic search.

Mitra et al (1998) have suggested a semantic search approach by applying query modification to the given input query. It plays an important role in many semantic search systems. Different techniques are developed to increase recall and precision of a query. The increase of precision is called query disambiguation. With ontology, it is easy to increase recall of a query using query-terms. Query modification is categorized as three way namely manual modification, query rewriting and graph-based approach. Query rewriting is subdivided as augmentation, trimming and substitution. In augmentation, query is enhanced with terms that are derived from ontological context of original query terms. Trimming of a query removes query-terms and has opposite effect of augmentation. In substitution, search terms are replaced with ontological related terms with synonyms and hyponyms from ontology to increase recall or precision.

Semantic search is achieved with conceptual graph using ontology. Conceptual graph is a diagrammatic representation graph which holds collection of logic forms. The use of ontology with structured semantic matching is one of the ways to improve precision and recall. There are some common methodologies available to implement semantic search approach which is RDF path traversal, keyword to concept mapping, graph patterns, logics, fuzzy concepts, relations and fuzzy logics.
RDF path traversal forms a group which contains arcs and paths to encode information. It is natural to apply semantic search. Query formulation is needed to search a path in RDF path traversal with ontological information for a particular domain. It is implemented by selecting classes and relationships which are used as constraints in actual search for instances. To implement keyword to concept mapping technique, the knowledge is formally encoded and it is an optimistic one. Mapping patterns in a graph sentence is used to give the relationships representation to the user and formulation of the user’s queries as natural language forms. Here, keywords are the entry points for locating information quickly.

Graph patterns are also used for implementing semantic search for getting answers for complex queries. RDF path traversal can be applied to narrow type of question and this graph pattern can only give an accurate result for complex queries. Logics are used for setting rules for retrieving the answer from ontology. Most inference algorithms support quick answer retrieval.

Kara et al (2012) have discussed an ontology based retrieval system with keyword based semantic retrieval approach. The performance of the system is improved using domain specific information extraction, inference and rules. Semantic indexing approach is used to improve scalability. There is a need to apply techniques for answering complex queries with few keywords. For this type, semantic querying is used. But, that type of querying language is needed a knowledge of the ontology and syntax of the language. Scalability of the system is improved in two way namely inference and querying. Inference is achieved by dividing the whole model into smaller individual model. Querying is achieved by using the technique is called inverted index structure that answers millions of
questions with reasonable time and retrieved answer from huge data source.

WSD algorithm is used to extract key concepts from a document collection. This algorithm is mainly used to remove stop words from a set of collection. In this system, WSD algorithm is used to improve the performance of retrieval and semantic indexing is also used instead for query construction from keywords. In semantic indexing, semantic data in knowledge base is indexed in structured way with keyword queries. This indexing is applied in knowledge base to index all the RDF triples. The naive indexing mechanism is used for extracting RDF triples. Typical approaches in this direction involve a combination of statistical technique like syntactic parsing and semantic operations to identify ontology concepts in user’s input.

Jones et al (2003) have analyzed the semantic web based QA system with effective information retrieval. In this system, a general approach for retrieving answer from the system is based on noun phrase identification, synonym set building, word sense selection, incorrect sense exclusion, hypernyms inclusion, inconsistency resolving, boolean query construction and retrieved the result. This system uses lexicons and ontology in effective manner for processing queries and also user intervention is low. But, the scalability and customizability is to be improved in this proposed model.

Wang et al (2007) have proposed automatic assistant learning mechanism for Electronic Learning (e-Learning) environment by analyzing learner’s question and find the relevant answer from domain ontology. A link grammar parser is used to analyze syntactic information from the given question. Depending upon the syntactic information,
queries are processed with similar word lists from WordNet by extending relevant meaning. Semantic tree is formed and the relevant contents are retrieved for answer. But, the pattern matching technique is to be improved for handling complex queries and automatic ontology learning is also being needed in their method.

Trillo et al (2007) have presented a semantic guide approach for discovering senses of set of user keywords. The main features of this approach are that it is an iterative approach to retrieve knowledge repositories in parallel manner and it is used to sense ontology terms in syntactic and synonym manner. This approach is also used to sense synonym between two senses by linguistic and structural similarity.

A new approach is discussed by (Guo 2008) which is used to analyze a question based on detection of question focus chunk, semantic chunk and question template. Vectors are used for semantic representation. A chinese natural language based system is implemented through NL parser. Semantic units and semantic rules are constructed for composing semantic primitives, senses and semantic chunks. The semantic primitive is the basic unit of hierarchy and it has all kinds of semantic features. Semantic chunks are used to express compound concepts and semantic rules are used as abstract of semantic knowledge. But, the domain ontology is expanded and ontology consistency is not ensured in this model.

Jiang & Tan (2009) have used a user modeling to capture user’s interest in domain ontology for providing personalized information service in semantic web. This model uses concepts, taxonomic relations and non-taxonomic relations for capturing user’s interest. A spread activation procedure is used for inference mechanism in user ontology. A statistical
method is used for learning user ontology. It is used as an integrated approach for both learning and exploiting user ontology. This model is integrated in semantic search engine and applied for retrieving document in digital library.

A challenged model of natural language QA with semantic search based web service is discussed by (Quarteroni 2011). A semantic of natural language is suggested for mapping semantics search with minimal human contribution and maximum domain coverage. A common domain diagram is created for services in knowledge base. But, efficient strategies are needed to acquire a service and effective answer extraction approaches are not suggested in this work. Alfonseca et al (2001) have developed a QA system with semantic distance. It is used to measure the module that to make use of all semantic relationships in WordNet for estimating semantic distance between question and answer.

Zhou et al (2007) have introduced a system called Adapting Keyword Query to Semantic Search (SPARK) system for adapting keyword query to semantic search approach. This approach is used to translate the keyword queries into formal logic queries by using semantic search automatically. It is used to provide Kruskal’s algorithm (Kruskal 1956) with indexing mechanism.

Kruskal’s algorithm is a greedy algorithm used in graph theory which is used to find a minimum spanning tree for a connected weighted graph. It means that it finds a subset of the edges that forms a tree which includes every vertex and the total weight of all the edges in the tree is minimized. This Kruskal’s algorithm is used to determine whether there is a valid ontology graph solution. In this thesis, all relations are allocated the same weight in RDF graph. If two concepts have a single relation,
then the weight is set as one. The weight of two concepts is set by depending upon the number of relations between them. External resource for indexing is used with ontology and KB. This indexing mechanism is applied to SPARQL query only. The query execution returns a set of tuples that satisfy the SPARQL query. Extract semantic entities from the tuples and access semantic index are used to collect all words in query which are annotated with these type of semantic entities. This semantic indexing is already discussed by Pustejovsky et al (1997) for hyperlinking process.

Semantic search with ontological QA is used a conceptual graph matching algorithm for improving precision and recall ratios. Several types of semantic search approaches are discussed and the graph pattern matching produces the correct answer retrieval for improving the precision and recall ratios. WSD algorithm is also used for extracting terms from documents. A parser with inference engine is used for implementing semantic search. Table 2.2 shows different semantic search approaches handled in different Ontological QA systems.

**Table 2.2 Comparison of Semantic Search Approach in Ontological QA Systems**

<table>
<thead>
<tr>
<th>Author Name</th>
<th>Approaches used</th>
<th>Performance metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mitra (1998)</td>
<td>Query modification in input query</td>
<td>Query disambiguation with precision and recall</td>
</tr>
<tr>
<td>Alfonseca (2001)</td>
<td>Semantic relationship</td>
<td>Semantic distance measurement</td>
</tr>
<tr>
<td>Quarteroni (2011)</td>
<td>Mapping semantic terms</td>
<td>Precision and recall measurement</td>
</tr>
<tr>
<td>Kara (2012)</td>
<td>Keyword based semantic retrieval</td>
<td>Inference and rules with semantic indexing</td>
</tr>
</tbody>
</table>
2.5 SURVEY OF SEMANTIC PATTERN RELATION CALCULATION

Hovy et al (2001) have discussed a semantic based analysis of question with candidate answers based on hierarchy of semantic patterns of question and answer. In this model, question answer pattern is generated automatically through learning process. A set of question and answer pairs are tied and matching criteria is used to identify the corresponding portion of tree. But, the exact syntactic and semantic information in answer model is lagging in this work.

Fleischman et al (2003) have discussed a QA system based on a repository of semantic relations between lexical terms extracted from text collection by POS tag patterns and a machine language classifier. In this model, information from web is collected offline and semantic relations are extracted. The repositories are automatically filled by a system which is used to watch online news sources and useful information. The patterns are generated manually. Each pattern has machine learned filter for achieving high precision.

Ferres et al (2004) have discussed a QA system which has a semantic representation of a question based on information from EuroNet and a set of semantic constraints. EuroNet is a global leader in processing electronic payment services. Answer extraction technique has an iterative relaxation of semantic constraints. It is used to support a Client/Server model which is managed by central server called metaserver. Metaserver is a central server which is used to collect data from different client in Client/Server environment. The advantages of this model are used to prevent from repeated initializations of components and it is used to allow easy debugging, maintenance and replacement of individual component.
Punyakanok et al (2004) have discussed a QA technique which is used to select an answer based on dependency tree matching. Question and answer are represented as dependency trees with semantic information and similarity is computed using approximate tree matching algorithm. A free form question answering task approach is implemented in this model. Both questions and candidate passages are represented using dependency trees augmented with semantic information. A generalized edit distance is computed between a candidate passage representation and the question representation. A distance is calculated in semantic similarity checking point of view. This distance is computed via a dynamic programming based approximate tree matching algorithm. But, this dynamic programming is used to calculate the cost of distance in tree mapping.

Sun et al (2005) design a syntactic dependency analysis for query expansion and the semantic analysis is based on frame based semantic representation generated by shallow semantic parser. This parse is used for semantic answer extraction. Answer passage selection is computed with frame similarity scores based on WordNet. In this model, dependency relation matching is proposed for query expansion. Dependency relation matching does not perform well in short questions with few keywords. So, additional context information for that short question is processed through query expansion. This query expansion algorithm can expand new terms and can find the relevant paths based on dependency relation analysis. With discussion about various semantic search based system, pattern matching is important one for answer retrieval which is discussed in following.
2.6 SURVEY OF RDF GRAPH MATCHING

Lu et al (2005) have introduced an approach to RDF query to manipulate and inference a database but query efficiency is improved by using inference mechanism. To reduce the cost of inference, batch mode and incremental mode type of inferences are used in this approach. RDFS Query and Manipulation Language (RQML) is introduced for easy and uniform data access in a declarative way.

McGlothlin & Khan (2009) have discussed the inference mechanism for representing RDF inference queries for knowledge management. In this model, inference rules are applied to RDF datasets for deriving additional facts through subsumption, symmetry and transitive closure. Inference is performed at storage time and inferred triples are persisted. Data is stored in relational database by using efficient schema model. Inference is performed in query as a simplification model and it is used to improve the performance. Inference involves with probabilities is lagging in this model. Harabagiu & Bejan (2005) have presented a QA methodology for handling temporal inference method based on relations of expected answer to temporal expressions in question or answers. This system is used to enhance recognition of exact answers to a variety of questions about time. Temporal information is required in different forms of inferences that derive from relations between events, their arguments, time available in discourse context. Annotation is generated based on this temporal phase only.

Kontos et al (2005) have discussed inference based QA system for biomedical domain with NL grammar for capturing casual reasons. Text analysis of relations is involved with casual knowledge. This model is suggested a method Automatic Representation Independent Syllogistic
Text Analysis (ARISTA) with ARISTA Oriented Modeling Adaptation (AROMA) system. It is used for intelligent text mining with knowledge discovery. Kontos et al (2005) also proposes the formulation of QA with casual knowledge expressed either linguistically or formally. A top-down approach for construction of NL grammars for QA is proposed. Grammar engineering methods such as hand crafting and machine learning is suggested. Katz et al (2002) have introduced a system named SynTactic Analysis using Reversible Transformations (START) in which natural language annotation is implemented. This system is used to analyze collections of natural language sentences and phrases which are used to describe the contents of information. A pointer gives the text segment with the answer sentence.

2.7 GENERAL PRINCIPLES OF AQUALOG QA SYSTEM

For the thesis work, AQUALOG QA system is taken for comparison. Compared to other types of QA system discussed in section 2.3, the performance of AQUALOG system is low but, the semantic search approach was implemented only in this system in question and answer extraction phase. AQUALOG is a portable question answering system which takes queries expressed in natural language and ontology as input and returns answer from the semantic web. AQUALOG system uses GATE NLP platform, string metric algorithms and WordNet. It is coupled with portable and contextualized learning mechanism which ensures the performance of system to improve over time.

AQUALOG is implemented in java using Client/Server environment. It is a plug in mechanism which allows different KR languages. Two main modules are available in AQUALOG. One is linguistic component and other one is relation similarity service.
Linguistic component is used to map NL input query to query triple set form. A set of syntactic annotations are associated with input query. These annotations are included in the information of English sentence, tokens, nouns and verbs. A set of annotations are achieved by using Java Annotation Patterns Engine (JAPE) grammar. JAPE grammar is a set of regular expressions using previous annotations in documents. Sometime, linguistic component cannot resolve ambiguity associated with NL query.

Relation similarity service component is invoked after NL query has been transformed into term relation form and classifies into approximate category. String distance metrics algorithm is used for relation term identification. Relation similarity service is designed for every step of user interaction. Learning mechanism is used with string metric algorithm and it is used to generate correct logical query. To avoid ambiguity, WordNet is used for finding synonym for particular term in query. When a question with similar context is asked, the relation similarity service cannot disambiguate the relation name.

2.8 PROBLEM STATEMENT

One particular challenging environment in closed domain question answering system with ontology and semantic web domain is the searching technique. Existing studies shows that the accurate answer retrieval from knowledge base is not achieved by using syntactic search. In this thesis, AQUALOG QA system is taken for analyzing the performance with precision and recall ratios. In this AQUALOG system, following problems are identified and these problems are solved with proposed techniques which are discussed in this thesis.
The first problem in AQUALOG QA system is the syntactic search. In user point of view, a same keyword gives different meaning. It is called ambiguity. This is the problem of getting correct result from AQUALOG QA system. Query triples with syntactic matching are used in this system. The triples are framed based on syntactic method.

The second problem in this QA system is the graph matching technique. Conceptual graph matching technique is used for query mapping which is also based on syntactic mapping. AQUALOG follows string metric algorithm based upon syntactic graph matching technique. So, the precision and recall ratios are low in this system.

The third problem is the syntactic indexing mechanism used in AQUALOG QA system. It does not produce correct answers for syntactic indexing mechanism. Answer extraction technique can be applied with graph based mapping scheme. So, the AQUALOG QA system is not suitable for improving precision and recall ratios.

2.9 OBJECTIVE OF THE PROPOSED WORK

For solving the above mentioned problems, this thesis proposed a question answering based on automatic learning system with ontology and semantic web in closed domain environment. This proposed QAAL system is designed to overcome the drawbacks of AQUALOG QA system for improving the performance of precision and recall ratios.

The first proposed method in this thesis is an improved question to query conversion method based on query template model. Mapping of users question and the query representation in ontology is a challenging task and this challenge is avoided by using a semantic search based query
template model in QAAL system. This query template model is created for implementing semantic search by improving precision and recall ratios. The input of Q2Q algorithm is the different types of question such as factoid based, list based, definition based, reason based and explanation based types and the output is the triple set form as subject, predicate and object format. This triple set is mainly used to represent RDF code for the next stage of QA. Question classification and reformulation takes a major role in this proposed Q2Q algorithm. Collections of SPARQL query templates are created as output of Q2Q algorithm and are referred after reformulation of query in user side.

The second proposed method in this thesis is the Question Answering based on Automatic Learning Graph Traversing (QGT) algorithm based on semantic similarity matching of RDF graphs. This algorithm is used to map query terms with knowledge base in ontology by using semantic similarity based graph matching technique. QGT graph matching is done with semantic similarity measurement among the concept, relations and graph in ontology. QGT matching with semantic similarity measure is used to improve the precision and recall ratios in QAAL system.

The third proposed method in QAAL system is the improved answer extraction mechanism by using semantic similarity based on integration of semantic indexing with inference mechanism. Inference is a method for knowledge acquisition and expansion. It extends knowledge base with additional information using metadata and rules. Semantic indexing is an index type which can make use of information extraction and annotation to index terms in ontology. Ambiguity has been avoided
and precision, recall ratios are high by using semantic search with semantic indexing mechanism with inference model. Confidence ratio is proposed to identify the resultant answer is correct or not.