DISSERTATION

On

Framework for automatic generation of answers to conceptual question in Frequently Asked Question (FAQ) based Question Answering System

Thesis Submitted in the partial fulfillment of the requirements for the degree of

Master of Technology

In

IT (Courseware Engineering)

Submitted By

Saurabh Pal
Examination Roll number: M69CWE10-01
University Registration Number: 103072 of 2007 - 2008

Under Guidance of
Assoc. Prof. Arindam Chakravorty
St Thomas’ College of Engineering & Technology

School of Education Technology
Faculty Council for UG and PG studies in
Engineering and Technology
JADAVPUR UNIVERSITY
KOLKATA- 700032
2010
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Executive Summary</td>
<td>7</td>
</tr>
<tr>
<td>Chapter 1: Introduction</td>
<td>9</td>
</tr>
<tr>
<td>Chapter 2: Literature Survey</td>
<td>13</td>
</tr>
<tr>
<td>Chapter 3: Problem Statement and objective</td>
<td>26</td>
</tr>
<tr>
<td>3.1: Problem Statement</td>
<td>26</td>
</tr>
<tr>
<td>3.2: Objective</td>
<td>26</td>
</tr>
<tr>
<td>3.3: Concept and Problem Analysis</td>
<td>26</td>
</tr>
<tr>
<td>3.4: Assumption and Scope</td>
<td>29</td>
</tr>
<tr>
<td>Chapter 4: Design / solution</td>
<td>31</td>
</tr>
<tr>
<td>4.1: Design</td>
<td>31</td>
</tr>
<tr>
<td>4.2: Implementation</td>
<td>37</td>
</tr>
<tr>
<td>Chapter 5: Results and interpretations</td>
<td>58</td>
</tr>
<tr>
<td>Chapter 6: Conclusions and Future Scope</td>
<td>74</td>
</tr>
<tr>
<td>6.1: Conclusion</td>
<td>74</td>
</tr>
<tr>
<td>6.2: Future Scope</td>
<td>76</td>
</tr>
<tr>
<td>References</td>
<td>77</td>
</tr>
<tr>
<td>Appendix I</td>
<td>80</td>
</tr>
<tr>
<td>Appendix II</td>
<td>128</td>
</tr>
<tr>
<td>Appendix III</td>
<td>136</td>
</tr>
<tr>
<td>Appendix IV</td>
<td>155</td>
</tr>
</tbody>
</table>
Executive Summary

Today’s web based or network based education demands for an Automatic Question Answering [QA] System which will automatically generates answer to user’s/student’s question. The QA System takes the students question as input in natural language and with many tools and techniques available in Natural Language Processing process the question. The answer is extracted either from a question and answer set [Frequently Asked Question Base] or from document set or by using knowledge base. The answer generated is sent back to the user.

In QA System the use of knowledge base is effective for factoid question and conceptual question but not found satisfactorily to answer the reasoning (why and how) type question, as answering these type of question needs to know the intension of question and synthesis of the answer using various concepts. Instead using Frequently Asked Question (FAQ) base gives a satisfying result, as answers corresponding to various forms of questions are available, including how, why, conditional etc. The input question is made matched with FAQ base question set, the semantically related question’s answer are extracted. The FAQ Base system fails if no semantically related question is found in FAQ base corresponding to the input question. Thus the obvious problem with FAQ base System is the preparation of Question and answer set, as it not known from previous what types and forms of question can be asked by a student. With different teaching and learning style different forms of question can come from student. This demands obviously an automatic answering system as backup, which will generates answers to the questions – especially those conceptual question (reasoning one) only if the FAQ based QA system fails. The question asked by the user/student may be out of scope or the question asked may not be required for the student at that level, also it is possible that the answer generated by the system may not be correct or the answer synthesized may not in the form which may be catered to student and feed into FAQ base, hence a
manual verification and authentication of question and it’s produced answers is required. This assures that the question and its answers in the FAQ base are correct.

The purpose of this research work is to design and develop a framework for automatic generation of answers to conceptual question especially “what”, “why”, “how” questions for Frequently Asked Question based Question Answering System, thus in term making a tool that will assist in generating answer to various conceptual question asked by the student and then helping in manual authentication of generated answers by an expert. The manual authentication verifies that the answer saved into the FAQ base is correct, and as per the student’s level of understanding and with in the scope and hence helps in slowly enriching the FAQ base for future queries. The present work applies the concept of semantic net to understand the semantics of the question with using less NLP task. The question focus is understood by understanding entities /concept present in the question and by the question word [“why, “how”, “what”]. The answer is obtained from the tagged document set and from knowledge base using semantic net and semantic frame. Information obtained from various concepts is joined together to make a whole total information for the answer. The answer produced along with other System generated messages helps the expert to understand where the system fails and how much more can be done to enrich the knowledge base.
1. Introduction

In today’s society with increasing education requirements the role of distance education become very important as it makes learning and training more accessible, convenient, effective and more cost efficient for learners and for education providers. Of the many ways of distance education, web based or network based learning has proved to be the most effective one. A web based or network based education follows a client- server architecture, where electronic material (a combination of different text, image, audio, video and other multimedia files) are uploaded into server which are accessed at client side the user/student from various geographical locations. While studying the content of the material many question may can rise in student’s mind. The questions asked by a student/user are generally either a factual question or a conceptual question. The factual question demands for names of entities, events that happened, date and other facts whereas the conceptual question asks for definition of terms, ideas, view points, classification, explanation, principles, relationships, reasoning – which explains how and why things. Answering the question raised is an important process of the e-learning system, as it improves and increases the efficiency of learning process. Through e-mail or chat a teacher or an expert can answer the student’s question, but this increases the teacher’s burden. This often demands for an Automatic Question Answering [QA] System which will automatically generate answer to user’s/student’s question. The QA System takes the students question as input in natural language and by using either a question and answer set or document set or by using knowledge base the answer is generated and sent back to the user.

The QA System takes the student’s question as input in natural language with many tools and techniques available in Natural Language Processing like word segmentation, stemming, Part Of Speech tagging, Named Entity Recognition etc, the Keywords and Named Entity are identified [3] which are then used in relevant document and passage searching and selection. Based on the question type the expected
answer type patterns [21][3] are generated which are then used for searching for a matching pattern in the extracted passage or in the relevant document. This process is effective in searching answer for factoid question. Pattern matching may not always produce satisfactory results for conceptual questions as the answer generated from required answer pattern often may not satisfies the intension or semantics of the question. For understanding the question, it need to be classified or categorized into what its asking for – name, place, object, date, event etc called Expected Answer Type[3] then using Named Entity Recognition technique the Named Entity is extracted from text, this process is effective only for factoid questions. For reasoning or conceptual question’s answering another approaches are used. Some of predefined question and answer pairs often called as Frequently Asked Question [FAQ] Base [22][10] is made prior to use. The input question’s semantic is understood and are matched with the available question in the FAQ base. Semantically related matching question’s answers are extracted and ranked. This process is effective than searching answer pattern for conceptual questions [10]. Presently more emphasis is given on the use of Knowledge Base, which includes Ontology – a conceptual map for understanding intend, semantics of input question and in searching answer. In QA System answer are extracted either from documents set or FAQ base or knowledge base [2]. To more recently a various combinations of FAQ base or document set or knowledge base is used for extracting answer. If the answer is not found in the FAQ base it is searched in the document set or generated automatically from the knowledge base.

Extracting answer from the document using keyword search, named entity recognition and answer pattern matching are found working for answering factoid /factual question in comparison to conceptual question. While retrieving answer from document, it is difficult to figure out where in the passage the answer is present and up to how much the answer would be – thus it becomes difficult to answer conceptual answer from the document set. Ontology is used as a tool to understand the semantics
of question by understanding the meaning and relation various concepts present in question as well as in text. This technique is effective for factoid question and conceptual question but not found satisfactorily to answer the reasoning (why and how) type question, as answering these type of question needs to know the intension of question and synthesis of the answer using various concepts.

Full extension use of NLP is often not satisfying to answer the conceptual (or reasoning) type question especially how and why type questions. Instead using FAQ base gives a satisfying result, as answers corresponding to various forms of questions are available, including how, why, conditional etc. The input question is made matched with FAQ base question set, the semantically related question’s answer are extracted. The FAQ Base system fails if no semantically related question is found in FAQ base corresponding to the input question. The obvious problem with FAQ base System is the preparation of Question and answer set, as it not known from previous what types and forms of question can be asked by a student. With different teaching and learning style different forms of question can come from student. This demands obviously an automatic answering system as backup, which will answer the conceptual question – especially those reasoning question if the FAQ based QA system fails.

A methodology for automatic QA System based on document tagging and knowledge base incorporating semantic net and semantic frame is provided for answering conceptual question’s answer – especially the reasoning ones as possible solution for making up a tool which help in understanding user’s/student’s question’s intension and answering it and thus in a process gradually helps in enriching the FAQ base.

The question asked by the user/student may be out of scope or the question asked may not be required for the student at that level, this requires a manual verification before the question and answer is put into FAQ base. It is possible that answer generated by the system may not be correct or the answer synthesized may not
in the form which may be catered to user/student and feed into FAQ base, hence a manual verification and authentication of question and it’s produced answers is required – so that the FAQ Base will contain correct and authenticated answers to question. The proposed system has provided with a mechanism that will assist a teacher/expert in understanding whether the answer is within the scope of knowledge base or not, if it’s not where it fails. The system makes it clear to what it had understood and what not and where it fails thus in terms helps in enriching the knowledge base for answering future queries.

The thesis is organized as:

- This section introduce about the background, problem and its solution.
- The second section contains the literature survey.
- Third section defines the problem statement, objective, concept and problem analysis, scope of work. This section explains the problem and what can be done more.
- Section Four describes the proposed methodology and implementation details.
- Section Five analyses the various output results.
- The sixth section contains a general discussion and what conclusion can be drawn and future scope.
2. Literature Survey

With the development of network technologies and multimedia distance education has become popular. Distance education has eliminated the restriction of traditional teaching [face to face mode teaching] on the time and space. The student can take lessons on its own convenience. The instruction materials are uploaded onto the internet which student can access, study and practice on his own convenience independently of teacher’s assistance. Often while studying or taking a lesson [instructions] various questions arises in students’ mind which needs to be answered. The student can get an answer to a question in different ways: By e-mail, chatting or by keyword search. The shortcoming of e-mail is that the e-mails are often not answered in time and answering e-mails for teacher is difficult when mails comes in large number every day. Chatting demands teacher to be remain online all the time where as the keyword search requires the knowledge of searching and it increases the student’s burden as the keyword search is unstructured and is difficult for student to articulate the correct keyword/s, which may possibly defocuses the student. Another option available is checking the Frequently Asked Question [FAQ] set from the FAQ base. A Frequently Asked Question [FAQ] base is a collection of different questions with their answers, which often people asked in a particular domain. The questions are indexed on some particular order for fast searching and retrieval. The question set is generally build by teacher or instructor, the obvious problem with FAQ base System is the preparation of Question and answer set, as it not known from previous what types and forms of question can be asked by a student. With different teaching and learning style different forms of question can come from student.

The Other alternative way could be Automatic Question Answering [QA] System based on Natural Language Processing for answering student’s question. Natural Language Processing (NLP) is a processing technique that encompasses a broad set of theories and techniques for automated generation, manipulation, and analysis of
natural or human languages. NLP converts Natural Language [speech or text] into a formal representation that is easy for computation. The Automatic Question Answering System takes questions from students in natural language and answers them using Ontology based knowledge base or FAQ base or documents or any combination of the resources. The Automatic QA System behaves as if the teacher is always present for answering the student’s query or in other words the system find out the intension of the student and subsequently produces the student’s need. Producing answer using automated QAS is difficult as it depends on the richness of the content or the way the question’s intend [semantics of question] is analyzed and understood.

The QA System can be either open domain or closed/restricted domain. QA System used in web based / network based teaching are generally closed/restricted domain. The QA System based on restricted domain uses document set, knowledge base or FAQ base or more often database to retrieve answers to input question from student. The document set is a collection of documents in textual format in a particular domain. Document used for information extraction can be structured or unstructured. A Knowledge Base is a store of facts, relationships and procedure that constitute the knowledge about a given domain represented in machine-processable form, In a Knowledge base the Knowledge acquired of a domain is represented using various knowledge representation techniques as rules, net, frames and scripts. Of the many knowledge representation techniques, Ontology is used for popularly. Ontology is a knowledge map, which describes the concepts and the various relations among concepts. Field/domain experts build Ontology based knowledge base either manually or by using semi automatic tool like Protégé. The different terms, concepts, attribute, relation and instance are extracted from corpus and are related to each other to make a concept model. Among other languages, Ontology Web Language [OWL] is one, which can be used to make all kind of ontologies. Based on the knowledge and its mapping the ontology can be of various types like domain ontology, vocabulary ontology, question mode ontology, application ontology etc. While obtaining answer to an input question
the QA System requires the understanding of what the question is, and retrieving the answer from the relevant documents having high probability of containing the answer or from the knowledge base or from FAQ base. The information extracted is often filtered to obtain appropriate answer to the question asked.

In open domain as the topic/subject on which the question will be asked is not known from previous, it is very difficult to define the set of documents required for answer extraction. For this very reason the open domain QA System uses the Internet [Uses search engines] to search and retrieve relevant documents with having high probability of containing answer. In restricted domain or closed domain the question asked are from fixed topics / subjects. As the topics / subjects are known from previous, the required documents and the knowledge base can be provided early to the system.

The main task associated with the QA System is, understanding the question imposed in natural language and extracting the meaning full answer from document/s, knowledge base, FAQ base, and some database. Question Answering (QA) involves the task of Natural language processing (NLP) and a type of information retrieval. All Question Answering System generally consists of three modules [21] Question Analysis Module, Search Module, and Answer Extraction Module.

The Question Analysis Module includes:

- Analyzes of question and determines its type.

- Understanding the semantics of the question.

- Determines the answer type, which will offer help for answer extraction.

- At the same time, it also translates natural language questions into queries for the search engine.
The Search Module retrieves the relevant documents or snippets or information using the queries generated by the first module. These retrieved results may contain answer to the question.

Answer Extraction Module retrieves the answer from knowledge base or FAQ base or extracts the answer from documents. While extracting answer from document sets it analyzes the documents or snippets or information retrieved from the second module, and then extracts the answer to the question.

While talking of answer extraction from document/s, it is found that answer extraction from an unstructured text/document is more difficult than from a structured text / document as in structured document the contents are kept in structured manner and are thus related in manner. An unstructured document can be made easily, where as the structured document has an overhead in building as it requires structuring and relating the contents [as data] with in the document for example XML files. Among the many techniques available – tagging is one used effectively for retrieving information from a unstructured document. Tag[12] are the user specified words used to semantically annotate the information/content which could be a word, phrase, sentence or passage in the document in order to describe, organize and correlate content with other contents[23] or simply to retrieve it easily. The limitation with tagging system is that it suffers from inconsistency, ambiguity problems, beside its very tedious job to manual tag a document content.

In the QA System all the methods, which are used to extract answer, are categorized into two classes – shallow method and deep method [11]. Shallow method uses keyword-based techniques to locate passages or sentences or information with probable chance of containing answer from the retrieved document. A keyword doesn't have to be a single word - sometimes it can be a phrase. In corpus linguistics a keyword is a word, which occurs in a text more often than normal. Deep method includes many techniques in itself namely Named Entity Recognition, relation detection, co –
reference resolution, word sense disambiguation, logical form transformation, logical inference and common sense reasoning or spatial reasoning. Deep methods are generally used for questions like why or How or dialog queries or question with deeper understanding.

It is found that that in its most simple form keyword or phrase search is best for retrieving relevant document from the document set or document base, irrespective of whether open domain QA system or closed domain QA system.

The keywords or phrase used for searching relevant document are extracted from the question. In an open domain QA System it is not known from previous which word is to be taken as keyword or which phrase are required. In restricted domain a domain lexicon or ontology is generally used to identify the keywords. In linguistics, the lexicon of a language can be defined as a vocabulary, containing all or all most all words and expression in the language in an alphabetical order. Beside these grammatical information [ Part Of Speech] of each word may also be present. A domain lexicon is an alphabetical list of terms in a particular domain of knowledge. Domain lexicon entries help to understand new or uncommon words and specialized terms present in question which otherwise may not be found in normal lexicon.

Different Natural Language Processing techniques are applied to segregate the keywords or the phrase from the question sentence such as:

Word segmentation – word segmentation is the process of segmenting different word used in a sentence.

Syntax analysis – Syntax analysis is the process of analyzing the input sentence to determine the subject, object etc

Stop word elimination – removal of un unwanted words like determinants, article, adverb etc,

Stemming – stemming is the process of obtaining the stem word or root word from a given word. The root word is a word, which is formed after removing all affixes
[suffixes and pre fixes]. -- e.g. "export" has a root word "port", sleepwalking has the root word "walk", and cannot be reduced into smaller constituents. A root word is the basic linguistic unit of a word or the basic part of a word that usually carries meaning.

Part of speech[POS] tagging[14] - In text word can be classified into one or more lexical or part-of-speech categories such as nouns, pronoun, verbs, adverb, adjectives, and articles etc. A POS tag is a symbol representing such a lexical category, e.g., NN (noun), VB (verb), JJ (adjective), AT (article). POS tagging is the process of automatically assigning POS tags to each word in the sentence.

The keywords thus are identified in a question sentence are then used to extract the sentences and passages from the relevant documents or some databases containing the answer, but then the sentence extracted may or may not have the required answer as the search is based on the keyword which is without understanding neither the meaning of the question and of the sentence/passage.

Beside keyword search answers from relevant document/s often found by using pattern matching [21]. The question or query is analyzed using NLP techniques to determine the question pattern and probable answer patterns, these answer patterns are then searched in the relevant document/s to obtain the answer. Considering a question “when was Abraham Lincoln born?” . The question pattern determined as “when be Q_PRN Q_DoVerb ?” Where Q_PRN is the question pronoun denotes “Abraham Lincoln” and Q_DoVerb is the verb denotes “born”. Based on this question pattern different answer pattern are generated like: Q_PRN(DATE), Q_PRN be Q_DoVerb in DATE etc. These answer patterns are then searched in the document set for matching patterns. This technique is used to obtain single line / word answer to factoid question.

Keyword techniques are not sufficient to understand the semantics of question and often much of the contextual processing is required, such as the Named Entity Recognition [NER][13][14] technique. Named Entity is a named object for Person, Organization, Location, Date and time, quantity (amount), a NER system uses a NE
Recognizer to recognize the Named Entity in a question sentence and hence classify these names into a set of predefined categories or class such as person, organization, location, date and time etc. The categories can be further specialized into subcategories. NER recognizes and select particular entity based on predefined category and thus it is typically used as an aid to filter input data and remove irrelevant data. NER are generally domain dependent. The required Named Entity [NE] is searched in the documents and in relevant sentences by an NE recognizer, the relevant word representing the required NE category are retrieved which are further filtered to obtain the answer based on what the question is looking for. The problems in NER are that they have semantic ambiguity. The other problem is making the taxonomy, categories and subcategories.

The NE recognition identifies and classifies the entities appear in natural language text be it a question or an answer. In case of question the NE only identifies the entity but that wouldn’t help in understanding what the question is asking for. An Expected Answer Type [EAT][3] or a question classification type is required which often determines what the question is looking for besides what the entities are present in the question. For example assuming a question “In which year did Einstein win the Physics Nobel Prize” the NER identifies “Einstein” as PERSON and “Physics Nobel Prize” as MISCELLANEOUS but it fails to recognize what the question is looking for - which is the “year”. EAT are predetermined categories or types – the question is searched for some pattern or words to understand the question type or the EAT. In the work done by Boldrini E. et al the NE and EAT of the question is identified, the EAT is the concepts which are looked in an ontology to identify exactly what is being required.

A combination of different techniques as the work done by Ashis kumar Saxena et. al [18] - keyword selection, noun phrase selection, question classification, Named Entity Recognition techniques is used to increase the efficiency of answer extraction from a set of documents. The question classification identifies and classifies what the
question is looking for. There are some predetermined patterns of question available based upon which it can be determined whether the question is looking for person, location, organization, number, title, jobtitle, date and money. The keywords and the noun phrase is used to retrieve document with high chance of containing answer. The NER is used to recognize the NE in the sentences extracted and hence the sentence with most matching NE with the question type is selected. In some the questions where its difficult to answer using NE, a pattern matching is done. Based on the question pattern an answer pattern is generated which is then matched in the sentences extracted from documents.

The Question Sentence beside containing nouns contains verbs , these verbs can be used to draw the relation between other parts of the sentence and thus the semantics of the question can be understood. The relation between the structure and the semantic meaning of verb can be deduced by Case grammar – a grammar theory proposed by Charles J. Fillmore. In a system proposed by Dunwei WEN et. al. [4] verb frames are detected to obtain the semantic information of question. The question sentence is syntactically analyzed by Treebank parser, which produces a tree structure of words and verbs used in the sentence. By using VerbNet (an instance of Case grammar) – hierarchical domain – independent, broad-coverage verb lexicon the verbs are classified and its Verb Frame are detected. The Verb Frame gives the syntactic, thematic role and its semantic information about the question verb. Question’s head verb frame is matched with verb frames of sentences in the retrieved documents, the sentence with matched verb frame is regarded as the answer.

The QA System based on the keyword, Noun Phrase, NE, Verb and pattern matching usually able to retrieve single word or single sentence factoid answers. The system often fails to generate relational, reasoning and logical answer. As these techniques of QA system uses mere words/terms or pattern for a match. These techniques are not able to determine the semantics of question or the question intension,
especially for complex conceptual question like reasoning, logical, conditional etc., there is a requirement to understand the semantics of question by understanding the relation and meaning of the concepts used to what the question is expecting.

The understanding of concepts used in question is one important process in question analysis, to very recent there has been a lot of work done on it. In restricted domain QA System, Ontology[17] is used a lot to explain the various concepts ,the various relations among them. Ontology is defined as an explicit formal specification of a shared conceptualization. It provides semantic meaning through relations between concepts. Ontology can describe concept accurately, especially the concept in special domain. Ontology, which is regarded as a modeling tools of conceptual model to describe information in the level of semantic and knowledge, can be applied to information and resources retrieval because of its good concept structure and ability of logic reasoning. The answer can be extracted straight from the Ontology or Ontology as a tool from the FAQ base, from retrieved documents/document set. Knowledge representation with ontology has proved to be a methodology to improve the semantic understanding capability of QA system. A QA System may can use one or more than one ontology. The use is mere depends upon the requirement and system design.

Xiaobo Wang et al.[20] used three ontology vocabulary ontology, question mode question mode ontology and domain mode ontology. The vocabulary ontology is used to refer the knowledge of different words and phrases used. The question mode ontology is used to identify the question mode. Field ontology describes the concepts and relationship among them. The words from question sentence are segmented using vocabulary ontology, using question mode ontology the main concept or the concepts and their attributes in the question sentence are identified, the relationship between concepts and attributes is obtained. . After the question mode is identified the answer is searched in the field/domain ontology.
Ontology is found very effective in solving complex conditional question, in the work done by Jie Lio [8] complex conditional questions are analyzed by using restricted domain ontology. The question are semantically analyzed to obtain question focus, question word [“what”, “why”, “where”, “how”, “when”] and the semantic chunk – the part of the question that has fixed semantic meaning like entities and attributes. It is found out whether the question is asking about entity, attribute, entity event, entity subclass, entity relation. The condition in the question and the question semantic makes the complex conditional question semantic representation. This representation is searched in the ontology together with the predefined rule for every condition type to extract the answer.

In the work done by Pengfei Zhang et al[15], ontology is used both to describe the resources and the domain. An auxiliary ontology is used to describe documents and concepts, where as the domain ontology is used to describe the relation between the various concepts in the domain. The question pattern is determined, the keywords in the question is identified by a domain vocabulary and the relation and meaning of the words are looked into domain ontology to understands the semantic of the question or what the question is asking for. The answer is searched in the domain ontology, if not found the documents described by the auxiliary ontology are searched. The text from the document corresponding to the matching concept as mentioned in question is extracted which then explains the concepts.

Wang Bo et al[19] uses different domain dependent ontology like top level ontology, domain ontology, task ontology and application ontology, these ontology are made prior either manually or by using semi automatic application like protégé. The question sentence is parsed and by process of word segmentation the words are segregated. The keywords extracted are mapped with the various ontology to understand the concept and the relation between the concepts. Using domain ontology the system can expands keywords, increasing the search area for the problem. The
system uses the expanded keywords to query the FAQ base, the answer of questions having most similar concepts is returned.

In QA System, use of FAQ base in comparison to document set is much more efficient for extracting answer, as the document set is generally useful only in extracting factoid and single line/sentence answer where as the FAQ base provides answer to various possible types of questions. Use of FAQ base for searching answer again has its own limitation, if no related question is found in the FAQ base – no answer will be selected, even though the concern concepts found in the ontology. To counter this generally a manually answering system or an automatic answer generation mechanism is used along with the FAQ base searching, in case the question searched for similarity is outside the FAQ base.

In the work done by Chun-Chia Wang et al.[2], a lexical parser is used to determine words, phrases and their POS in the input question. The keywords are identified and using WordNet [by carrying the Word Sense Disambiguation], each of the keywords synset-id is determined which represents the unique id of a concept or sense. Using direct Boolean match of sense-id of user question’s keyword and those of questions in Question Answer Set are matched. All the matching question’s answers are retrieved and ranked. A list of answer provided by the answer generator is displayed to the student/user. If no satisfactory answers are found the user send back the question to the System which in turn sends the question to the teacher/expert.

In the work done by Jiban Fu et al. they had used two approaches to obtain answer [6]. In one approach the question sentence is processed with the help of an Ontology, the question is classified into a predefined class [what’s the question’s intention is] and the concepts present in question are extracted. Questions with having similar class and concept are searched in the FAQ base. The most similar question’s answers are extracted. In another approach, if the user’s questions don’t match in the
FAQ base, the answer is extracted in support of ontology and logic reasoning. Similar to this in another model [17] given by Qinglin Guo, the user question semantic comprehension model is made by combining many Natural Language Processing tools, including Ontology, Segmentation, Part – Of – Speech Tagging, determining question type, keyword extraction etc. The most similar semantic question is selected from the FAQ base. If the answer to the question is not found in FAQ base, relevant document is fetched from the document base and using expected answer pattern for named entity information the relevant passage is extracted from the document.

In all kind of Question analysis and answer searching there is a draw back of understanding the English words as the same word may mean different things - polysemy and different English word may mean the same thing - synonymy. Identifying the meaning of the word in the very context is Word Sense Disambiguation. To disambiguate or identifying the meaning of the word or any other words having same meaning, machine readable lexicon – WordNet [5] is used. Each word from context question are looked in WordNet for similar senses. The similar meaning / concepts of two distinct words are combined in every possible pair combination to make different queries which are then searched in the documents or Knowledge Base.

QA Systems developed addresses factoid answers like single word, single sentence or retrieving answer from FAQ base based on the question semantic similarity. Qinghua Zheng et al [16] developed a system uses the knowledge oriented answer extraction method to answer questions beyond the factoids. In this system the unstructured document is analyzed by a lexical parser, using the techniques of segmentation and speech tagging the text is divided into sentences and passages. The sentences and passages retrieved is classified the semantic meaning of each individual is understood. The passages and sentences are then combined to generate the answer unit.
In web based/network teaching, the automatic question answering system takes the question from user/student in natural language and generates the answer (collection of answers). A list of answer provided by the answer generator is displayed to the student/user. If no satisfactory answer are found the user send back the question to the System which in turn send the question to the teacher/ expert[1][2]. The teacher/expert answers the question and sends to the user/student and correspondingly adds the new question and its answer to the FAQ Base. The benefit of the system is that day by day the new questions are added to the Question and Answer set and thus enriching the FAQ base.

In another approach the system understands the question focus or semantic and then the results are extracted automatically from knowledge base. If the system fails to generate the answer, the system gives the question to the teacher / expert. The teacher gives the answer to the question raised by the student and correspondingly manages the knowledge base [7]. In this way the non answerable questions of system are answered and the knowledge base is updated from time to time so that any question of similar type can be answered in future.
3. Problem Statement and objective

3.1 Problem Statement

To design a framework for automatic generation of answers to conceptual question in Frequently Asked Question [FAQ] based Question Answering System.

3.2 Objectives

To make a tool that will generate answers of different frequently asked questions raised by different users and helping in manual authentication of generated answers by an expert.

3.3 Concept and Problem Analysis

Use of FAQ base in QA System gives satisfying results, as answers corresponding to various forms of factual and conceptual question are available. The difficulty with FAQ base QA System is making of question answer set or the FAQ base in prior, as it not known from previously what questions would be asked by the student/user. The FAQ database starts with some standard questions and then eventually grows as the students/user make new queries. The input question is made matched with FAQ base question set and semantically related question’s answers are extracted. FAQ based QA system often fails if no semantically related question is found in the FAQ base corresponding to the input question. If the system unable to produce answer to the question, it is sent back to be answered manually [2]. The question is manually answered by an expert and sent back to the user. The new question and its answer is saved and indexed [22] into the Frequently Asked Question database. This process brings burdens over the teacher/domain expert as has to manually answer the unanswered question.
To overcome this QA System based on FAQ base are often combined with auto answer generation system, which generates answer to unanswered question from document set or from knowledge base [17]. The answer obtained is then sent back to the user. The System creates a cycle of question generation, answering and feedback [2][9]. The student raises the question, the answer is extracted from the Knowledge base and sent back to the student and saved in the FAQ base.

The problem with this situation is that the answer generated may be wrong, redundant, or the answer generated may not satisfy the semantics of the question properly. The wrong or improper answer may often lead to confusion in student. This requires for a mechanism where the answer are need to be authenticated for correctness before it can be sent back to the student or saved in the FAQ base.

Different techniques are used in, extracting answer from document set or knowledge base. While searching in a document set, the relevant documents are extracted by keyword, the keyword used are taken from the input question. The keyword search particularly identifies and locates the passage and sentence in the document, but it fails to say which word/phrase or how much text/sentence will give the answer. This generally requires the understanding of question semantics that is what the question is asking for – name, object, date, time, measurement, event, location etc and
is said as Named Entity. This determines the answer type and probable answer patterns. The generated answer pattern / template [14] are searched in the extracted sentence, passage or paragraph obtained by keyword search or often the expected Named Entity is searched by NER in the extracted passage and document and filtering out relevant String and discarding irrelevant data. This system is very effective in answering factoid/factual question answer but fails to answer the conceptual question answer (definition, reasoning, difference etc). The answer to conceptual or complex question can be single passage or multi sequenced passage. As it is not known which sentence, passage and text contains the information and it is thus difficult to answer the conceptual question from document set.

In the work done by Qinghua Zheng et. al [16] in the complex type question processing, passages are extracted, the semantic of the passages are understood and are combined to generate the answer. This system would fail where different concepts are involved as answer need to be built by understanding the concepts involved in the question, their relations and the given condition of the question. Understanding of the concepts used in question is one important task in understanding question meaning. For understanding concepts used in the question often a knowledge map – semantic net, ontology is used. In the work done by Boldrini E. et al [3], the question expected answer type is determined, which is often the concept used in the ontology, the instance of the concept is produced as the answer to the question. This system is only helpful in answering factoid question.

In the work done by Pengfei Zhang et al[15], the question concept is verified from ontology, after understanding the semantics the answer is extracted from propositional knowledge base in RDF/OWL model based on ontology. If the answer is not found it is searched in the document library. In this system pre-information are feed into the RDF about the description of each concept. This system fails to synthesize
answer where there are multiple concepts in question as this system works on one concept.

In the work done Xiabo Wang et al. [20], they have resolved problem of many concepts relation. It contains various concepts description and list of all attribute its value. The system fails to artificially synthesize answer and deliver only the stored answer based on the correctness of the relation between the concepts in the ontology thus fails to answer the reasoning based question answer as it need synthesis of answer from various concepts. It also fails to describe why attribute have that value.

The above discussion can be summed up for the following requirements

1. Requirement of a mechanism for generating answer to various conceptual question asked by user/student and hence in turn help in building FAQ base. This in turn demands for:

   1.1 A framework for extracting conceptual information from document set like definition, explanation, short note, contrast, feature, process, reasoning etc.

   1.2 A framework for extracting conceptual information from knowledge base and synthesis of answer.

2. A mechanism required for authenticating the answer generated by the QA System.

3.4 Assumption and Scope

- The QA System is tested, analyzed and verified on a Technical domain.
- Questions of what, why and how can be answered, so the question sentence need to be start with either “what” or “why” or “how” word.
• The question sentence should contain relevant keyword as present in the subject matter, which will help to understand the semantics of the question and can navigate the system to search the answer.
• Use of short forms, alphanumeric, numbers are restricted.
• The questions need to be done after reading the subject
• It is assumed that the same question or similar question is not found in the FAQ Base.
4. Design / Solution

4.1 Design

Technical domain questions generally contain concepts and attributes. The semantics of the question can be well understood by the presence of concepts and their attributes. The concepts are the entities about which the question is all about and the attributes define what is being needed to be asked/known about the concepts. For example in the question “why semiconductor have valency 4”, the concept present in it is “semiconductor”- which defines that the question is about semiconductor and the attribute “valency” defines what is need to be asked or get to be known about the concept “semiconductor”. The concepts in the question are the keywords, which are of course the Nouns. The concepts are the entities (classes and object) that are represented by a knowledge/concept map which contains all the concepts used in the domain and the inter relationship among them. Every concept has its description and own set of attributes. The attributes are noun like “resistor”, “resistance” and verb like “resistvity”. Thus noun, verb found in question is assumed to be an attribute which need to searched against the attribute set of concepts.

While answering a general conceptual question, explanation of the concept/s in question is necessary. Question having more than one concept, where one concept has an action on other concepts many in between interdependent concepts need to be explained. A conceptual question (“why” & “how” type) is of the form – “Why / How ....W……Z ?”, where W and Z are concepts, the question can have only one or more than one concepts. It is found that one concept has an action against another concept or one concept is cause/reason and the other concept is its result. To explain this kind of conceptual question its very important to get the relation between the concepts, like W – rel1 – X – rel2 – Y – rel3 – Z The relationship states that to explain one concept which are the interdependent concepts need to be explained. To get relation among the
various concepts in a domain a conceptual map is required which defines the various concepts in the domain and how they are related and connected to each other. For the question having attribute along with the concepts like “why / how….X….a…Y ?” where ‘X’ and ‘Y’ are the concept and ‘a’ is assumed to be an attribute of any present concept in the question. The attribute defines what need to be asked or get to be known about the concepts. Assuming ‘a’ may be an attribute of any concept present in question – searching the attribute in the attribute set of the concepts in question or in the concepts present in between the relationship of ‘X’ and ‘Y’ for a description regarding the attribute and its value can give the answer to the question. As it is not known the attribute ‘a’ belongs to which concepts, the answering to the question “why / how….X….a…Y ?” thus requires searching the attribute ‘a’ in both the concepts as well as in concepts present in the relationship between ‘X’ and ‘Y’. The matched attribute’s description is obtained from the concepts and would provide the answer to the question.

A Technical Questions may consists of :

- One concept with no attribute
- One concept with one attribute
- One concepts and many attribute
- More than one concept with no attribute
- More than one concept with one attributes
- More than one concept and attributes

Using a semantic net, concepts map is made. In a domain the semantic net defines how one concept relates with another concept. It is a directed graph consisting of vertices/nodes, which represent concepts, and edges, which represent relations between concepts. There is no standard set of relations for semantic networks, but the following relations are very common:
**INSTANCE**: X is an INSTANCE of Y if X is a specific example of the general concept Y. Example: Mohan is an INSTANCE of Human

**ISA**: X ISA Y if X is a subset of the more general concept Y. Example: sparrow ISA bird

**HAS - PART**: X HAS - PART Y if the concept Y is a part of the concept X. (Or this can be any other property) Example: sparrow HAS tail

Figure 1. Shows a section of semantic net used to represent the domain “semiconductor physics”, for detail refer Appendix III – Semantic Net

Every concept has its own set of attribute, which can described by semantic frame. The semantic frame consists of slots, each slot contains fields called facets which specify attributes, attributes value and description or comment. The description is a brief note on why the attribute has that value or an explanation. Below is a semantic frame of
For every concept found in question its searched and located in semantic net, and in a way relation among the concepts is derived, this would explain in what way the concepts are related. The interdependency among the concepts in question is obtained
from the relationship derived. The attributes found in the question are searched against each of the concept in the relationship.

For the Question having only one concept like “what is semiconductor?” with concept “semiconductor”. The concept is searched in semantic net and the corresponding semantic frame representing the concept is found out. The description of the concept is obtained from the semantic frame, which provides the answer to the question.

For the question having one concept and one attribute like “why semiconductor have valency 4” has got one concept “semiconductor” and one attribute “valency”. The concept is searched in the semantic net as well as the corresponding semantic frame representing the concept is found out. The attribute is searched in each of the slots both in the attribute field and attribute value field for a match. The matching slots description is extracted as the answer to the question.

In question with having more than one concepts and no attribute, the concepts are located in the semantic net and the relation between them is obtained. Description of each concept in the relationship is obtained from the corresponding semantic frame. The information thus obtained is combined together to form the answer.

In question with having more than one concepts and attributes, the concepts are located in the semantic net and the relation between them is obtained. Each of the attributes is searched for a match in the semantic frame for each concepts identified in the question as well as in the concepts in between their relationship. For every no match found for the attribute in the semantic frame, the description of the semantic frame is obtained and presented as information. The information thus obtained is combined together to make an answer to the question.
In another approach to answer conceptual question of “what” type, the information are extracted from text document using tag word. A technical subject-domain document content and its information can be categorized into different sections and parts as introduction, definition, brief, explanation, types, property, features, behaviour, function, procedure, components, parts, structure, necessity, application, theorem, principle, formula, example, figure, conclusion etc. Based on this different tags are made like:

1. <introduction>…………….</introduction>
2. <def>……………….</def>
3. <brief>……………...</brief>
4. <explanation>………… ..</explanation>
5. <types>…………………..</types>
6. <property>……………….</property>
7. <feature>…………………….</feature>
8. <behaviour>…………..…….</behaviour>
9. <function>……………..……</function>
10. <procedure>……………….</procedure>
11. <components>……………..</components>
12. <parts>……………………..</parts>
13. <structure>…………………</structure>
14. <necessity>…………………</necessity>
15. <application>………………</application>
16. <purpose> …………………</purpose>
Each tag has attribute name, of with keyword (concept name) as value. The tag should have at least one attribute in it. The attribute name defines what is being tagged and of define of whom. For detail on tagged document on “semiconductor” subject used in the program refer to Appendix III – tagged document.

While processing a “what” question, tag words or its synonyms are searched in the question, if found the concepts present in the question are searched with the tag name and tag’s attribute value as the concept’s name in the tagged document set. The matched tagged information is extracted as answer.

4.2 Implementation

The whole System is developed into three parts – i) Question processing, ii) Answer generation, iii) Answer authentication

The Question processing process takes the question sentence in text from user written in Natural Language, assuming the same or semantically related question is not found in the FAQ base – the question is processed. The question processing is done to correct spelling, word trimming, Base word conversion, removal of words with least
significance (words stemming), extraction of concepts and attributes. The question processing is done in following steps:

1. The input question is segmented into tokens (words, symbols etc) using space as delimiter.
2. Each token is checked in lexicon for spelling mistake. For each word not found in the lexicon is considered as spelling mistake.
3. Based on the first character match with having 60% same character count similar words are enlisted from domain vocabulary as possible correct spelling options.
4. The user is prompted to enter the correct spelling. The user can give the correct spelling, in case the word input is considered to be correct one, the user may continue with same.
5. Words are checked against the words in domain vocabulary and are converted to base/root word. In domain vocabulary for each base/root word the corresponding synonyms, words with spelling variation, plural forms, verb forms are kept. For example the input word “electrons” is converted to its root word “electron”. Similarly synonyms are mapped to a root word.
6. Each word is checked in lexicon, if found the word is tagged with appropriate part of speech (POS) like noun, pronoun, verb, adverb, modal verb, adjective, article etc.
7. Each word is again checked in domain vocabulary, if found the word is tagged with “keyword”.
8. From the first word of the input question sentence the type of question is determined. For the word is “what” – the question is definite and for “How”, “Why” the question is reasoning.
9. All other words are removed except for the words tagged with “NOUN”, “VERB” and “KEYWORD”. Thus a normalized question is obtained with only concepts and attributes.
10. Searching each word in the question into the semantic net identifies the concepts in the question. For a word found in the semantic net is considered as a concept.
of the domain and added to a nodes array and for the word not found in the semantic net are assumed to be an attribute which need to searched in the attribute set of concepts and hence added to property array.

The **Answer generation** process takes the nodes array and property array as input from the **Question processing** process. The Answer generation process obtains the answer either from the document base or from the knowledge base – consisting of semantic net and semantic frame. Based on the first word the question sentence the question type is determined that is whether its “what”, “why” or “how” question. The “why” and “how” types of question are put into reasoning based question. For “what” type of question the system looks for the tag word in the question, for the tag word is found the answer is extracted from the document set else it is obtained from knowledge base. For “why” and “How” type of question the information from the knowledge base is obtained and combined together to form the answer.

The Answer generation process is done in following steps:

1. First word of the question is checked for question type whether “what”, “why” or “how”.
2. For the question is “what” type, the tag word is searched in the normalized question. If the tag word is found each element in node array is searched with tag word as pair in the document set. For reference to tagged document set used in this system, refer to Appendix III.
3. If the pair (tag word and node array element) is found, the tagged information is obtained else, for the pair not found in the document set the knowledge base is searched for the answer.
4. If the question contains no tag word then the knowledge base (semantic net and semantic frame) is searched for answer.
5. For “what” question with no tag word found and for “why” and “how” questions the relation between the nodes in the nodes array is obtained from the semantic net. For reference to semantic net used in this system, refer to Appendix III.
6. Each of the property array elements are searched in the semantic frame of each concept in the nodes array for a match against attribute or attribute values. Matching slot description is extracted as information to the answer. For reference to semantic frame refer to Appendix III.

7. The various information obtained are joined/combined to synthesize the answer to the conceptual question.

8. The answer generated is stored in the database.

The **Answer authentication** process takes the unauthenticated answer from the database and sent it to the teacher/domain expert. The expert authenticates the answers produced by the system. The process saves the question and its authenticated answer into the database as well as sent the answers to the corresponding user/student who have raised those questions.
4.3 Flow Chart

Start

Input Question in Natural Language

Segment the sentence into words

Check spelling mistake for each word

Spelling mistake found?

Yes

Display correct spelling options

No

Input correct options

Prepare the question sentence with correct spelling words
Segment the sentence into words

Tag each word according to Part of Speech using lexicon

Tag the words for keywords using domain vocabulary

Remove all the words in the question except Noun, Verb and Keyword to form Normalized Question

For each word in the normalized question search the semantic net

Is found?

No

Yes

1

2

3
1. Add the word to Nodes array
2. Add the word to Property Array

Do all the words in the normalized Question are searched?

Yes

- Is Nodes Array is empty?
  - Yes: Display “The question lacks information”
  - No: Check the first word of question for Question type
    - Is it “what”? 
      - Yes: 1
      - No: 2

No

- Check the first word of question for Question type
Search for the tag word in the normalized question

Is Tag word found?

Search the tag word with each element of nodes array in the tagged document set

Is the tag word and concept pair found?

Extract the tagged information and store in an answer array

Display "tagged information not found"

3

4
Is it "why" or "How"?

Yes

Relate each word in the nodes array using Semantic Net. Store the relation in an answer array

For each element in the nodes array
Search the corresponding Semantic frame

For each element in the property array search for a match in the semantic frame’s slots

Is a match found?

No

Yes

Store the description of the slot in the answer array

Store the description of semantic frame in an answer array
Answer is checked for duplicate information.

Is duplicate information found?

Remove duplicate information from answer array

Display information from answer array

Stop
4.4 Algorithm

Step 1: The question is taken as input in the form of string

Step 2: The question is segmented into individual words using white space as delimiter.

Step 3: Each segmented word is matched in the domain vocabulary. If the word does not find a match in the lexicon, its set as wrong spelling.

Step 4: If each word spelling is found correct then go to step 8.

Step 5: For each wrong spelling founds a corresponding matching spelling with same first character and having 60% same character count is obtained from the domain lexicon.

Step 6: Words with its wrong spelling and the correct spellings are displayed as possible solution of the word to the user for input.

Step 7: Correct spelling words are taken as input from the user as a replacement for the wrong spelling words.

Step 8: Each segmented word is matched in the domain vocabulary, for a match found its converted to its base word/root word.

Step 9: The words that are found in the domain vocabulary are identified and tagged as “KEYWORD”.

Step 10: Using a lexicon, part of speech for other words beside keywords in question are identified and tagged [e.g., noun pronoun, verb, adverb, adjective, modal verb etc]

Step 11: A normalized question is obtained by eliminating pronoun, auxiliary modal verb, adverb, adjective, article from the question and selecting only keywords, noun and verb.

Step 12: If the first word is “what”. Set the question type to “what_type” else go to step 18.

Step 13: The normalized question is searched for tag words.

Step 14: If the tag word is not present, go to step 18.

Step 15: If tag word is present, the tag word is searched in the tagged text file/document against the keywords present in the normalized question.
Step 16: If the tag word and the keyword pair is found, the text between the starting and ending tag is displayed, else display “The tagged information corresponding to the keyword not found.”

Step 17: If the first word is “why”, “how”, set the question type to “why_type”.

Step 18: Each word of the normalized question is matched with every node of the semantic net.

Step 19: If found equal, the name is added to the nodes array else the name is added to the property array.

Step 20: Search is performed on the semantic net and semantic frames using the elements of nodes array and property array. Call search (nodes array, property array).

Step 21: The duplicate information in the information array is removed.

Step 22: The information stored in the information array is displayed.

Algorithm for SEARCH (nodes array, property array)
Step 1: for node array length = 1 and property array length = 0, call SEARCH1(nodes array[0]).
Step 2: for node array length =1 and property array length >0,
for each element in the property array
   flag = call search2(nodes array[0],property array[i]).
   if flag = false ,
   display “Since property array found no match in the properties of nodes array, the description of will probably give the result”
Step 3: call SEARCH1(nodes array[0]).
Step 4: for node array length >1 and property array length >= 0
       call SEARCH3(nodes array, property array).
Step 5: For node array length = 0 , display “The keywords / words not found in the knowledge base”.
Algorithm for SEARCH1(element): searching single node in the semantic net
Step 1: The element is searched against every node in the semantic net.
Step 2: If the element searched is found, the description of the node from the corresponding semantic frame is retrieved and stored in information array else set flag to false.

Algorithm for SEARCH2(element, property): searching single concept with single property in the semantic net
Step 1: The element is searched against every node in the semantic net.
Step 2: If the element searched is found, the property is matched for each attribute in the corresponding semantic frame.
Step 3: If the property gets matched with attribute or its value in the frame, the description of the attribute is retrieved and stored in an information array else the property is searched in base nodes or in the nodes which are one step up in hierarchy to the current node
Step 4: If base node found, go to step 2, else set flag to false.

Algorithm for SEARCH3 (nodes array, property array) : searching multiple nodes with multiple property in the semantic net
Step 1: The elements are searched against every node in the semantic net.
Step 2: The routes between the nodes are mapped.
Step 3: If the route do not exists, set flag to false.
Step 4: if flag = true , Relation between the adjacent nodes in the route is stored in an information array.
Step 5: For each node in the route/path the corresponding semantic frame is searched for each element of the property array.
For each node in the route
{
  For each element in the property array
call SEARCH2(node array[i], property array[i])
}
If SEARCH2(node array[i], property array[i]) did not found a match call
SEARCH1(node array[i])
}
Step 6: If flag = false, display “Route between the nodes don’t exists”

4.5 Case based illustration of algorithm

1. Question: what is semiconductor
Step 1: The question “what is semiconductor” is taken input in the form of string.
Step 2: The question is segmented into individual words using white space as delimiter.
On segmenting the question into words, three words are obtained:
“what”, “is”, “semiconductor”
Step 3: Each segmented word – “what”, “is”, “semiconductor”, is matched in the
lexicon. As match found they are considered as correct spelling.
Step 8,9: Each of the segmented word is compared or matched with in the domain
vocabulary. Matched word is tagged/identified as “KEYWORD”. Here the word
“semiconductor” found a match in the domain lexicon and hence tagged as
“KEYWORD”
“What”
“is”
“semiconductor” → KEYWORD
Step 10: Other words beside the word tagged as “KEYWORD are searched in lexicon
for a match, part of speech of the words are obtained and tagged [e.g., noun pronoun,
verb, adverb, adjective, modal verb etc]. The word “what” is located and identified as
Question Pronoun and “is” as To Be verb. They are tagged as:
“what” → QPRON
“is” → TOBE

Step 11: Normalized question is obtained by eliminating “what” and “is”.
Normalized question: “semiconductor” → KEYWORD

Step 12,13: The question type is set to “what_type”. Tag words are searched in the normalized question “semiconductor”.

Step 14,18: As in the question no tag word is present, the keyword “semiconductor” is searched against every node in the semantic net.

Step 19: As the keyword “semiconductor” is found in the semantic net, it is added to nodes array.

nodes array: [“semiconductor”]

property array: []

Step 20: search method is called with arguments nodes array and property array. Since the nodes array length = 1 and the property array length = 0, search1 method is called with nodes array[0] (“semiconductor”) as argument. The search1 finds the semantic frame(named “semiconductor”) corresponding to the nodes array[0] element and extracts the description field information and stored in the information array.

Information array = [“semiconductor: semiconductor is an electrically neutral substance, with no free electrons or holes. Its neither insulator nor conductor. Its resistivity lies somewhere between the both. Semiconductor conductivity depends much on temperature. On increasing the temperature the resistivity decreases making it conductive e.g., germanium, silicon”]

Step 21,22: The information array is checked for duplicate and the information in the array is displayed as the probable answer to the question.

Probable Answer: semiconductor: semiconductor is an electrically neutral substance, with no free electrons or holes. Its neither insulator nor conductor. Its resistivity lies somewhere between the both. Semiconductor conductivity depends much on temperature. On increasing
the temperature the resistivity decreases making it conductive e.g., germanium, silicon

2. Question: what are the features of co-valent bond

Step 1: The question “what are the features of co-valent bond” is taken input in the form of string.

Step 2: The question is segmented into individual words using white space as delimiter. On segmenting the question into words, 7 words are obtained:
“what”, “are”, “the”, “features”, “of”, “co-valent”, “bond”.

Step 3: Each segmented word – “what”, “are”, “the”, “features”, “of”, “co-valent”, “bond”, is matched in the lexicon. As match found they are considered as correct spelling.

Step 8,9: Each of the segmented word is compared or matched with in the domain lexicon. Matched word is tagged/identified as “KEYWORD”. Here the word “co-valent” and “bond” found a match in the domain lexicon and hence tagged as “KEYWORD”.

“what” → QPRON
“are” → TOBE
“the” → ARTICLE
“features” → VERB
“of” → QPRON
“co-valent” → KEYWORD
“bond” → KEYWORD

Step 10: Other words beside the word tagged as “KEYWORD are searched in lexicon for a match, part of speech are obtained and tagged [e.g., noun pronoun, verb, adverb, adjective, modal verb etc]. The word “what” is located and identified as Question Pronoun, “are” as To Be verb, “the” as article, “feature” as verb and “of” a preposition, They are tagged as:
“what” → QPRON
“are” → TOBE
“the” → ARTICLE
Step 11: Normalized question is obtained by eliminating “what”, “are”, ”the”, “of” as they are question pronoun, to be verb, article, preposition respectively and selecting keyword – “co-valent”, “bond” and verb – “features”.
Normalized question : “features” → verb, “co-valent” → KEYWORD, “bond” → KEYWORD

Step 12,13: The question type is set to “what_type”. Tag words are searched in the normalized question “features”, “co-valent”, “bond”

Step 15,16: As the question contains tag word “features”, the tagged text document is searched for tag named “feature” for each attribute “co-valent” and “bond”.
The search is found against the “feature” and “co-valent” pair the information is extracted and displayed.

probable answer:

i) Co-valent bonds are formed by sharing of valence electrons
ii) In the formation of co-valent bond, each valence electron of an atom forms direct bond with the valence electron of an adjacent atom. In other words, valence electron are associated with particular atoms. for this reason, valence electrons in a semiconductor are not free.

3. Question: why germanium have crystal structure

Step 1: The question “why germanium have crystal structure” is taken input in the form of string.
Step 2: The question is segmented into individual words using white space as delimiter. 
On segmenting the question into words, 5 words are obtained:
“why”, “germanium”, “have”, “crystal”, “structure”. 

Step 3: Each segmented word – “what”, “are”, “the”, “features”, “of”, “covalent”, “bond”, is matched in the lexicon. As match found they are considered as correct spelling.

Step 9: Each of the segmented word is compared or matched with in the domain lexicon. Matched word is tagged/identified as “KEYWORD”. Here the words “germanium” and “crystal” found a match in the domain lexicon and hence tagged as “KEYWORD”

“Why”
“Germanium” → KEYWORD
“have”
“crystal” → KEYWORD
“structure”

Step 10: Other words beside the word tagged as “KEYWORD are searched in lexicon for a match, part of speech of the words are obtained and tagged [e.g., noun pronoun, verb, adverb, adjective, modal verb, conjunction etc]. The word “why” is located and identified as conjunction, structure as “verb” and “have” as verb. They are tagged as:

“Why” → CONJUNCTION
“have” → MODAL VERB
“structure” → NOUN

Step 11: Normalized question is obtained by eliminating “why”, “have” thus
Normalized question is made up of “germanium” → KEYWORD, “crystal” → KEYWORD, “structure” → NOUN.

Step 17,18: The question type is set to “why_type”. Each word in the normalized question is matched with every words in the semantic net.

Step 19: The keyword “germanium” found a match against a node in the semantic net, and hence added to nodes array, where as the keyword - “crystal” and noun - “structure” did not found a match in the semantic net and hence added to property array.

nodes array: [“germanium”]
property array : [“crystal”, “structure”]

Step 20: search method is called with arguments nodes array and property array. Since the nodes array length = 1 and the property array length > 0, method search2(nodes array[0], property array[i]) is called for every element of property array.

The search2 method searches the semantic frame(named “germanium”) for the attribute “crystal” and “structure”. The attributes are searched in the current node and upon not founding are searched in the semantic frames of the base nodes. information is obtained from the slot of semantic frame where the attribute is matched and stored in an information array.

Search1(node array[0]) is called, which identifies a semantic frame named germanium its description is stored in the information array.

Step 21,22: The information in the information array is displayed as probable answer.

Probable Answer : In semiconductors all the atoms or molecules are arranged in orderly pattern and thus forms crystal shape

germanium : germanium is a semiconductor

4. Question: why impurity is added to semiconductor

Step 1: The question “why impurity is added to semiconductor” is taken input in the form of string.

Step 2: The question is segmented into individual words using white space as delimiter. On segmenting the question into words, 6 words are obtained:

“why”, “impurity”, “is”, “added”, “to”, “semiconductor”.

Step 3: Each segmented word – “why”, “impurity”, “is”, “added”, “to”, “semiconductor”, is matched in the lexicon. As match found they are considered as correct spelling.

Step 8,9: Each of the segmented word is compared or matched with in the domain vocabulary. Matched word is tagged/identified as “KEYWORD”. Here the words
“impurity” and “semiconductor” found a match in the domain lexicon and hence tagged as “KEYWORD”

“Why”
“impurity” → KEYWORD
“is”
“added”
“to”
“semiconductor” → KEYWORD

Step 10: Other words beside the word tagged as “KEYWORD are searched in lexicon for a match, part of speech of the words are obtained and tagged [e.g., noun pronoun, verb, adverb, adjective, modal verb, conjunction etc]. The word “why” is located and identified as conjunction, structure as “verb” and “have” as verb. They are tagged as:

“Why” → CONJUNCTION
“impurity” → KEYWORD
“is” → TO BE
“added” → VERB
“to” → PREPOSITION
“semiconductor” → KEYWORD

Step 11: Normalized question is obtained by eliminating “why”, “is”, “to” thus

Normalized question is made up of
“impurity” → KEYWORD
“added” → VERB
“semiconductor” → KEYWORD

Step 17,18: The question type is set to “why_type”. Each word in the normalized question is matched with every words in the semantic net.

Step 19: The keywords “impurity” and “semiconductor” found match against nodes in the semantic net, and hence added to nodes array, where as the verb - “added” did not found a match in the semantic net and hence added to property array.

nodes array : [“impurity”, “semiconductor”]
property array : [“added”]
Step 20: search method is called with arguments nodes array and property array. Since the nodes array length > 1 and the property array length > 0, search3(nodes array, property array) method is called.

search3 method traces the route between node “impurity” and “semiconductor”, and stores the relation between adjacent nodes in the route in an information array.

For each node in the route, corresponding semantic frame is searched for each element of the property array by calling search2 method.

If the search2 method fails to find a matching attribute in the semantic frame of the current node or base nodes, search1 method is called with nodes array element is passed as argument.

Step 21,22: The duplicate information in the information array is eliminated and displayed as the probable answer.

**Probable Answer:**
Extrinsic semiconductor is - a semiconductor
Extrinsic semiconductor has impurity

Semiconductor : semiconductor is an electrically neutral substance, with no free electrons or holes. Its neither insulator nor conductor. Its resistivity lies somewhere between the both. Semiconductor conductivity depends much on temperature. On increasing the temperature the resistivity decreases making it conductive

extrinsic semiconductor : extrinsic semiconductor are intrinsic semiconductor with impurity added

impurity : Impurity are element with have valencies > 4 or < 4. Intrinsic semiconductor has little current conduction, to increase its current conduction suitable impurity is added to intrinsic semiconductor. The purpose adding impurity is to increase its conductivity by increasing either the number of free electron or holes, this causes the increase in current conduction.
5. Results and Interpretations

An instruction material on semiconductor physics – is used as learning material for student, for more refer to Appendix III – learning material on semiconductor. Based on the study material, different question are posed. The question are made strictly using “what”, “why” and “how” as the initial word of the question sentence.

Samples of different “what”, “why”, “how” - input questions with the generated answers are shown below:

Result I

1. what are the properties of semiconductor ?

Generated Answer:

Properties of Semiconductors

i) The resistivity of semiconductor is less than an insulator but more than a conductor

ii) Semiconductor have a negative temperature co-efficient of resistance i.e. the resistance of a semiconductor decreases with the increase in temperature and vice versa. for example, germanium is actually an insulator at low temperatures but it become a good conductor at high temperatures.

iii) when a suitable metallic impurity (e.g., arsenic, gallium etc.) is added to semiconductor, its current conducting properties changes appreciably. This property is most important

2. what is a crystal

Generated Answer:

Crystal : A substance in which the atoms or molecules are arranged in an orderly pattern is known as a crystal.
3. what is the difference between intrinsic and extrinsic semiconductor

Generated Answer:

intrinsic semiconductor is-a semiconductor
extrinsic semiconductor is-a semiconductor

Since "difference" found no match in the properties of "intrinsic semiconductor" , "extrinsic semiconductor" , "semiconductor" , the description of "intrinsic semiconductor" "extrinsic semiconductor" , "semiconductor" will probably give the result

Intrinsic semiconductor : A semiconductor is an extremely pure form is known as an intrinsic semiconductor. The number of free electrons and holes created in intrinsic semiconductor are equal in number, hence the electrical conduction of intrinsic semiconductor is neutral.

Extrinsic semiconductor : Extrinsic semiconductor are intrinsic semiconductor with impurity added

4. what purpose impurity is added to semiconductor

Generated Answer:

The process of adding impurities to a semiconductor is known as doping. The amount and type of such impurities have to be closely controlled during the preparation of extrinsic semiconductor. Generally, for 10 to power 8 atoms of semiconductor, one impurity atom is added.

The use of adding impurity is to increase either the number of free electrons or holes in the semiconductor crystal. If a pentavalent impurity (having 5 valence electrons) is added to the semiconductor, a large number of free electrons are produced in the semiconductor. On the other hand, addition of
trivalent impurity (having 3 valence electrons) creates a large number of holes in the semiconductor crystal.

5. why semiconductor have co-valent bond

Generated Answer:

Since Semiconductor have valency 4, it cannot lose electron or gain extra electron to make ionic bond it thus share electrons with the neighboring atoms to form co-valent bond to complete its octet

6. why silicon have crystal structure

Generated Answer:

In semiconductors all the atoms or molecules are arranged in orderly pattern and thus forms crystal shape

silicon : silicon is a semiconductor

7. why silicon have valency 4

Generated Answer:

In semiconductor the outer shell of atom have only 4 electron and require 4 more electron to complete its octet

Since "valency" found no match in the properties of "silicon", the description of "silicon" will probably give the result

silicon : Silicon is a semiconductor. Silicon is an element in most of the common rocks,. Actually, sand is silicon dioxide. the silicon compounds are chemically reduced to silicon which is 100% pure for use as a semiconductor.

8. why intrinsic semiconductor is electrically neutral

Generated Answer:
The number of free electrons and holes created in intrinsic semiconductor are equal in number, hence the electrical conduction of intrinsic semiconductor is neutral.

In an intrinsic semiconductor, even at room temperature, hole current pairs are created. When electric field is applied across an intrinsic semiconductor, the current conduction takes place by two processes, namely by free electrons and holes. The free electrons are produced due to the breaking up of some covalent bonds by thermal energy. At the same time, holes are created in the covalent bonds. Under the influence of electric field, conduction through the semiconductor is the sum of currents due to the flow of free electron towards the positive field and holes to the negative field, but since the total number of electron and holes created are equal in number the electrical conduction is neutral.

9. How semiconductor forms the bond

Generated Answer:

Since "forms" found no match in the properties of "semiconductor", the description of "semiconductor" will probably give the result.

semiconductor : semiconductor is an electrically neutral substance, with no free electrons or holes. It neither insulator nor conductor. Its resistivity lies somewhere between the both. Semiconductor conductivity depends much on temperature. On increasing the temperature the resistivity decreases making it conductive e.g., germanium, silicon.

Since Semiconductor have valency 4, it cannot lose electron or gain extra electron to make ionic bond it thus share electrons with the neighboring atoms to form co-valent bond to complete its octet.

10. How temperature effects the conductivity in semiconductor

Generated Answer:

semiconductor depends-on temperature
The property effects, conductivity, found no match in the properties of temperature
The property effects, conductivity, found no match in the properties of semiconductor

temperature: The conductivity of semiconductor depends on temperature. When the temperature is arised, some of the covalent bonds in the semiconductor break due to the thermal energy supplied. The breaking of bonds sets those electron free which are engaged in the formation of these bonds. The result is that a few free electrons exist in the semiconductor. These free electrons can constitute a tiny electric current if the potential difference is applied across the semiconductor. Thus the conductivity of semiconductor increase with rise of temperature.

At absolute zero temperature, all electrons are tightly held by the semiconductor atoms. The inner orbit electrons are bound whereas the valence electrons are engaged in co-valent binding. At this temperature, the covalent bonds are very strong and there are no free electrons, hence the semiconductor crystal are insulator at absolute zero or conductivity decreases with the decrease in temperature.

Probable Answer: semiconductor: semiconductor is an electrically neutral substance, with no free electrons or holes. Its neither insulator nor conductor. Its resistivity lies somewhere between the both. Semiconductor conductivity depends much on temperature. On increasing the temperature the resistivity decreases making it conductive e.g., germanium, silicon

11. How hole current is created and effects the conductivity of semiconductor

Generated Answer:

p-type semiconductor has hole current
p-type semiconductor is a extrinsic semiconductor
extrinsic semiconductor is a semiconductor
The property created, effects, conductivity, found no match in the properties of hole current.

The property created, effects, conductivity, found no match in the properties of p-type semiconductor.

The property created, effects, conductivity, found no match in the properties of extrinsic semiconductor.

The property created, effects, conductivity, found no match in the properties of semiconductor.

Hole current: At room temperature, some of the co-valent bonds in pure semiconductor break, setting up free electrons. Under the influence of electric field, these free electrons constitute electric current. At the same time another current - the hole current - also flows in the semiconductor. As the current flow is the presence of holes in the co-valent bonds. Therefore, it is more appropriate to consider the current as the movement of holes.

P-type semiconductor: When a small amount of trivalent impurity [elements like gallium or indium] is added to a pure semiconductor, it is known as p-type semiconductor. The trivalent impurity increases the number of holes in the semiconductor.

12. How impurity increases the electrical conductivity of semiconductor.

Generated Answer:

Extrinsic semiconductor has impurity.
extrinsic semiconductor is-a semiconductor

The property increases, conductivity, found no match in the properties of impurity

The property increases, conductivity, found no match in the properties of extrinsic semiconductor

The property increases, conductivity, found no match in the properties of semiconductor

impurity: Impurity are element with have valencies > 4 or < 4. Intrinsic semiconductor has little current conduction, to increase its current conduction suitable impurity is added to intrinsic semiconductor. The purpose adding impurity is to increase its conductivity by increasing either the number of free electron or holes, this causes the increase in current conduction. The impurity can be trivalent [valence electron = 3] or pentavalent [valence electron = 5].

extrinsic semiconductor: extrinsic semiconductor are intrinsic semiconductor with impurity added

semiconductor: semiconductor is an electrically neutral substance, with no free electrons or holes. Its neither insulator nor conductor. Its resistivity lies somewhere between the both. Semiconductor conductivity depends much on temperature. On increasing the temperature the resistivity decreases making it conductive. e.g., germanium, silicon.

For more questions on “what”, ”why” and “how” refer to Appendix IV- Section 1.

The questions are posed to the system with concepts which will found in the knowledge base, the answer generated by the system are as per assumption made according to the algorithm. To understand how the system would behave with unknown concept, unrelated concepts and concept present in different net, wrong questions are given to the system. Few questions with their respective generated messages are listed below, for more refer to Appendix IV – Section II.

1. how p-type semiconductor is used in pn junction
2. what is the valency of carbon

Generated Message:

*** The keywords / words not found in the knowledge base ***

3. what are the feature of doping

Generated Message:

*** The tagged information corresponding to the keyword not found ***

4. how indium is useful in increasing conductivity

Generated Message:

*** The keywords / words not found in the knowledge base ***

5. how impurity causes excess electrons

Generated Message:

Since "causes" found no match in the properties of "impurity", the description of "impurity" will probably give the result

Since "excess" found no match in the properties of "impurity", the description of "impurity" will probably give the result

Since "electrons" found no match in the properties of "impurity", the description of "impurity" will probably give the result

Probable Answer: Impurity: Impurity are element with have valencies > 4 or < 4. Intrinsic semiconductor has little current conduction, to increase its current conduction suitable impurity is added to intrinsic semiconductor. The purpose adding impurity is to increase its conductivity by increasing
either the number of free electron or holes, this causes the increase in current conduction. The impurity can be trivalent [valence electron = 3] gallium (At. No. 31) and indium (At. No. 49) or pentavalent [valence electron = 5] arsenic (At. No. 33), and antimony (At. No. 51).

Interpretation I

After analyzing the answer of the question’s given it can be said that the:

- The program can generate answer in most of the cases by using tagging and through use of semantic net.
- The program can identify if the keywords or concept are found or not in the knowledge base
- The program also can found whether there exists any relation between concepts.
- The program can also found out long relation between the concepts present in the question.
- The program suitably produces the message, which enables the expert to actually understand where and how the program fails to produce answer.

After testing with many variations of question it is found that the program fails to process the input question sentence properly under some condition, the possible solution of it is provided to the program to enhance the question processing efficiency.

- The program fails to resolve synonyms [words with similar meaning], thus Keywords with similar meaning are treated as different words.

  Keeping similar or synonym words in a table can solve the problem. So each matching word can be found.

- The program goes for exact word match while searching for keywords in domain lexicon, semantic net and in semantic frame. This often causes failure in search as often the typed keyword in question may vary in writing [spelling] with the
keywords in vocabulary or in the semantic net or in semantic frame. For example, user may write the word “pn-junction” whereas in system’s vocabulary it’s in the form “pn junction” and hence the word “pn-junction” and “pn junction” are treated as different words causing word mismatch and failure in search. Similarly, words “electron” and “electrons”, “resistance” and “resistivity” are treated as words with different meaning.

The problem can be solved by keeping relative similar word in a table. For each match found in the table the root word is extracted.

- The program fails to correct the spelling mistake – That misspelled words are not detected and corrected.

This problem can be solved by detecting the incorrect spelling, and giving the most similar words as option for correct spelling to the user.

- Abbreviations or short forms are often considered as incorrect spelling.

This problem can be solved by detecting the incorrect spelling, and giving the most similar words as option for correct spelling to the user and then prompting the user to continue with same or the given options.

- The program fails to correctly generate correct answer if in relation between two concepts, if one concept is present as attribute or attribute value of other concept as well as another node in the semantic net.

The concepts are checked for has- a relation in each other, if the concept is found to be an attribute to another concept it is not searched in the semantic net.
The tag word list may be incomplete and thus proper tagging of all or various kind of information in the document may not be possible.

After updating the program, questions with wrong spelling, synonyms etc are feed into the program. The program processes the question with the above amendment. Sample question with their respective generated messages and interaction with the user are listed below, for more refer to Appendix IV – section III

Result II

Q. what is semcondtr

Question Processing :

The input question :what is semcondtr
Word Segmented : what is semcondtr
Wrong word :semcondtr
Correct word may be :semiconductor,
Enter the correct word [Enter to skip]: semiconductor
Corrected Spelling :semiconductor
The question consists of :3 words
Question After trimming /word conversion: what is semiconductor

Q. what are the properties of sican

Question Processing :

The input question :what are the properties of sican
Word Segmented : what are the properties of sican
Wrong word :sican
Correct word may be: semiconductor, silicon,

Enter the correct word [Enter to skip]: silicon

Corrected Spelling: silicon

The question consists of: 6 words

Question After trimming /word conversion: what are the properties of silicon

Q. how semiconductor depends on temperature

Question Processing:

The input question: how semiconductor depends on temperature

Word Segmented: how semiconductor depends on temperature

Wrong word: temperature

Correct word may be: trivalent impurity, temperature,

Enter the correct word [Enter to skip]: temperature

Corrected Spelling: temperature

The question consists of: 5 words

Question After trimming /word conversion: how semiconductor depends on temperature

Q. what is n-type material

Question Processing:

The input question: what is n-type material

Word Segmented: what is n-type material
Wrong word : \textit{n-type}

Correct word may be : \textit{n-type semiconductor, n-type conductivity,}

Enter the correct word [Enter to skip]:

Corrected Spelling : \textit{n-type}

The question consists of : 4 words

Question After trimming /word conversion: what is \textit{n-type semiconductor}

\textbf{Interpretation II}

- The program can solve synonyms [words with similar meaning].
- The program can detect incorrect spellings and provides correct spelling option to user.
- While drawing relation between concepts, if concepts present at both, node position in semantic net as well as attribute or attribute value in semantic frame of other concepts in question, the program can resolve which concepts to be searched in semantic net and which concepts need to be searched as attributes in other concepts frame.

\textbf{Result III}

After analyzing the answer generated it is found that the answer contains many extra, unrelated and unimportant information. Based on the relevance of information contained in the answer the , the answer are categorized into 4 categories.
Category “A”: The answer contains exact pieces of information,

Category “B”: The answer contains exact piece of information with addition to that 1 or 2 pieces of extra information.

Category “C”: This answer contains exact piece of information as well as many other pieces (>2) of extra information.

Based on the above metrics the answer generated for “what”, “why” and “how” are evaluated. For 20 questions each on “what”, “why” and “how” run on the program the following results is obtained.

1. Answer to “what” type question

Number of Category “A” answer: 17
Number of Category “B” answer: 3

2. Answer to “Why” type question

Number of Category “A” answer: 13
Number of Category “B” answer: 6
Number of Category “C” answer: 1

3. Answer to “Why” type question

Number of Category “A” answer: 11
Number of Category “B” answer: 8
Number of Category “C” answer: 1
Interpretation III

There are various reasons, which may result in generating answers with extra information along with the correct information piece.

- The question input may have more than one concept. As the distance between the concepts corresponding to the nodes in the semantic net increases, unwanted or extra information may come in. The algorithm obtains the route or the relation between the concepts corresponding to the nodes in the semantic net, if the number of in-between nodes increases the information/description of each node along with the correct piece of information is obtained and displayed. These extra pieces of information – description about all in-between nodes may often not require.

- The extra information increases for each unmatched word (verb or noun) in the question. The words are matched against attribute or attribute’ value in the frames of each concept in the relation obtained between the concepts present in

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>What</td>
<td>17</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Why</td>
<td>10</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>How</td>
<td>7</td>
<td>10</td>
<td>3</td>
</tr>
</tbody>
</table>
the question. For every non-match the description of the frame is obtained and displayed.

- For the concepts in the questions are not related, the description of the concepts along with the description matching attribute or attribute’s value is given. So the answer contains the exact information along with extra description of the concepts which may often not need.
6. Conclusions and Future Scope

6.1 Conclusion

The program generates answer based on the concept present in the question and can be used as tool for generating answer to Frequently Asked Question raised by student. Use of knowledge base (semantic net and semantic frame) and tagged document set is very helpful.

The document tagging system is quite effective in tagging or semantically annotating the content. The tag set used in the system may not be complete and requires more tags for other type of domain. The text/passage extraction from document is a definite problem as it is not known for sure what the text/sentences means, the semantics of the passages or sentences were not known this lead to the problem of extracting answers for conceptual question. The use of tag has solved the problem to an extent as it annotates the text or sentences or passage with meaningful information and thus helps in extracting the required piece of information. The limitation with tagging system is that no standards for tags are available.

For understanding the meaning of question use of concept is quite effective and its requires much less Natural Language Processing tasks. The use of knowledge map is definite requirement as this technique is the very effective in understanding the various concepts used in the conceptual question. Construction of proper conceptual map is very critical and requires domain expertise. Improper or incorrect concept may lead to total failure of answering system.

The program can relate concepts, concepts with attributes etc. it produces most of the answer to the questions having various relations and attributes in it. The answer produced contains statements of various relations among concepts with description of concept and the description of any reasoning found in the semantic frame. The answers produced often are not in a format to be catered directly and thus requires a manual editing and authentication. The program produces answer with the assumption and
philosophy that producing the relation among the required concepts and description of the relation of attribute with concept, if no relation is found with concept then mere description of the concept will obviously covers lot of the answer part in it. It is found after testing with different test cases that for exact tag match in case of “what” question and for “why”/“how” questions where attributes founds a relation with concepts in the knowledge base- semantic frame in most of the cases where attributes don’t found any relation with concepts or the distance between the concepts in the semantic net is more or use of more number of concepts in question, it generally fails to cater the exact answer. The answer obtained contains either exact answer or information about the relation among the concepts, description of attributes found in relation with concepts and descriptions of required concepts. The information thus obtained contains either the exact answer or exact answer with other extra information or the whole information thus obtained together makes the answer. The information obtained needs an expert for manual editing and authentication before suitably catering it to student or save it in database as FAQ set. The program generates message for every instance when i) It don’t found the required tag for a concept in the document ii) Concepts are not found in the concept map [semantic net] iii) Concepts are not related in the concept map [semantic net] iv) Properties in question don’t match against the attributes or its value in the semantic frame of required concept.

The messages produced by the program will helps a lot to the expert to change and upgrade the knowledge base [semantic net and semantic frame] and adding tags and other required information to the document so for future quires the program can suitable produce the right full answer.

The program generates the answer without taking consideration of the knowledge level of student. A question’s answer can be given in many ways and forms based on the knowledge level of the student. Answering a student’s question needs understanding of what concepts need to be included and how much the concept should be described based on the knowledge level of student. The program generates answers
based on its knowledge base (semantic net, semantic frame) and document set irrespective of whom the answer needs to be delivered. It thus becomes the job of expert to authenticate and edit the answers produced by the program while understanding for whom the answer meant to be.

The program using the concept of knowledge base – the semantic net and semantic frame produces information against an input question, but requires a manual authentication for validating the information as answer. Thus the framework can be used as tool for generating answers to Frequently Asked Questions and thus in a way can help in enriching the FAQ base.

6.2 Future Scope

- The program is tested only using concepts and concepts relation, without doing much of the NLP tasks. A Syntax analysis can be done to determine which the primary concept is and which the secondary concept is and so on. This would determine which concept depends on which one.
- The tags used are on a single technical subject and needs more to be tested on other domain, so that tag set can be made complete and a standard on tag can be obtained.
- The program is tested with only one level of semantic net and may need to be tested with different levels from semantic net from abstract to detailed level. This might help in reducing the extra information obtained in answer while deducing the relation between the concepts.
- The program is tested in one technical domain and may need to be tested in other technical domain for effectiveness.
References


