A Framework For Automatic Generation Of Answers To Conceptual Questions In Frequently Asked Question (FAQ) Based Question Answering System

Saurabh Pal
Dept of CSE and IT,
Bengal Institute Of Technology,

Kolkata, India. Sudipta Bhattacharya
Dept of CSE and IT,
Bengal Institute Of Technology,
Kolkata, India.

Indrani Datta
Kolkata, India

Arindam Chakravorty
Dept of IT,
St Thomas College of Engineering & Technology,
Kolkata, India.

***Abstract*- Question Answering System [QAS] generates answer to various questions imposed by users. The QAS uses documents or knowledge base for extracting the answers to factoid questions and conceptual questions. Use of Frequently Asked Question (FAQ) base gives a satisfying results to QAS, but the limitation with FAQ base system is in the preparation of Question and Answer set as most of the questions are not predetermined.QAS using FAQ base fails if no semantically related questions are found in base, corresponding to the input question. This demands an automatic answering system as backup, which generates answer especially to conceptual questions.The work presented here gives a framework for automatic generation of answers to conceptual question especially "what", "why" and "how" type for frequently asked based, Question Answering System and in terms gradually builds up the FAQ base.**

Keywords- Question Answering System; Frequently Asked Question; Frequently Asked Question base; knowledge Base; semantic net; semantic frame; tagging.

I. INTRODUCTION

In today's society with increasing information requirements the web based Question Answering System [QAS] is needed. The question asked by user is either a factual question or a conceptual question. The factual question demands for name of entity, events that had happened, date and other facts whereas the conceptual question asks for definition of terms, ideas, views, classification, explanation, principles, relationships, reasoning – which explains how and why things. This demand for an automatic QAS which will generate answer to user's input question. The QAS takes the question in natural language as input and by using a question - answer set or document set or knowledge base generates the answer.

The QAS takes the user's question as input in natural language and identifies keywords, Named Entity[2] present in the question by using techniques available in Natural Language Processing like word segmentation, stemming, Part Of Speech tagging, Named Entity Recognition etc. The identified Keywords and Named Entity are then used for searching and selecting relevant documents, passages from the document set. Based on the question type and pattern the expected answer patterns (templates) [6][2] are generated which are subsequently used for searching a matching pattern in the extracted passage or in the relevant document. As generated answer patterns, often may not satisfy the intension or semantic of question, understanding the semantic of question is required. The input question is classified or categorized into what it's asking for – name, place, object, date, event etc called Expected Answer Type [2], and then by using Named Entity Recognition [NER] technique the expected Named Entity is extracted from the text (document). Extracting answer from documents using keyword search, answer pattern matching and named entity recognition founds effective for factoid / factual questions then in comparison to conceptual questions. The difficulty in extracting answer to conceptual question from document is in figuring out, where in the passage the answer is present and up to how much the answer would be. For answering the reasoning or conceptual questions other approaches are followed, such as use of Frequently Asked Question [FAQ] Base [7], Knowledge Base [1]. FAQ Base is a set of predefined question and answer pair made prior to use. Use of FAQ base involves understanding the semantic of input question and matching it against the available questions in the question and answer base. The semantically related matching question's answers are retrieved and ranked. This process is effective than searching answer patterns for conceptual questions [3]. Presently more emphasis is given on the use of Knowledge Base, which includes Ontology – a conceptual map used as a tool for understanding the intension/semantics of input question by knowing the meaning and relation of various concepts or entities present in question as well as in the relevant text document for extracting the answer. To more recently various combinations of FAQ base, document set and knowledge base is used for extracting the 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. These

techniques are not found fully satisfying to answer the reasoning (why and how) type questions, as answering these types 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 type questions especially those how and why types. Instead using FAQ base gives a satisfying result, as answers to various forms of questions are available in the FAQ base, including how, why, conditional etc. The problem associated with FAQ base System is in the preparation of Question and answer set, as it not known from previous what type and forms of question will be asked by users. This demands obviously an automatic answering system as backup, which will answer to conceptual question – not found in FAQ base.

A framework for automatic QAS based on tagged document and knowledge base incorporating the semantic net and semantic frame is provided here for answering conceptual [especially reasoning ones] question's answer –as a possible solution for making up a tool which will help in understanding the question's intension and answering it and thus in a process gradually helping in enriching the FAQ base.

The proposed framework is provided with a mechanism that makes clear to what it had understood and what not and thus where it fails. This in terms helps in enriching the knowledge base by assisting the expert, in filling the gaps within the knowledge base so that future queries may be answered properly.

II. Problem Defination

Use of FAQ base in QAS gives satisfying results, as answers corresponding to various forms of factual and conceptual questions are available. The difficulty with FAQ base is in building question and answer set in prior, as it is not known from previously what questions would be asked by users. The FAQ base starts with some standard questions and then eventually grows as users make new queries.

The input question is matched with FAQ base question set and semantically related question's answers are extracted. The system fails if no semantically related questions corresponding to the input question are found in the FAQ base. The unanswered question is then sent back to be answered manually [1] later by an expert. The new question and its answer are then updated in FAQ base [7]. As the unanswered questions could be many, the process will bring burden over the domain expert. To overcome this QAS using FAQ base are often combined with auto answer generation system which generates answer to unanswered question from document set or from knowledge base [5].

With various NLP complexities, it is possible that the auto answer generation system may generate wrong answers or have redundant information in it, or the answer generated may not satisfy the semantics of the question properly. The wrong or improper answer may often lead to confusion in user. This requires for a mechanism where the answers are authenticated for correctness before it can be sent back to the user and saved in the FAQ base.

Generating answer automatically requires extraction of information from document / knowledge 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 using keywords, obtained from the input question. The keyword search particularly identifies and locates the passages and sentences in the document, but it fails to identify which word/phrase or how much text/sentence will make the answer. This generally requires knowing what Named Entity the question is asking for such as name, object, date, time, measurement, event, location etc. The expected Named Entity is searched by NER in the extracted passage and filtering out relevant String and discarding irrelevant data. The Named Entity by this also determines the answer type and probable answer patterns. The generated answer pattern / template [4] are searched in the extracted sentence, passage or paragraph obtained from keyword search to generate the answer. These techniques are very much effective in answering factoid/factual question's answer but fail to answer the conceptual question (definition, reasoning, explanation, difference etc). The answer to conceptual or complex question can be single passage or multi sequenced passages. Due to the natural language complexities it is not identifiable which sentences, passages and how much text contains the information and is thus difficult to answer the conceptual question from document set. The other problem with reasoning based questions with multiple concepts[8] or entities present in it, is in understanding the relation between the concepts, and there after artificially synthesis of answer with different concepts.

The above discussion can be summed up for the following requirements

1) A question's answering mechanism for generating answer to various conceptual questions raised in FAQ based QAS and hence in turn helps in building FAQ base. This in turn demands for :

a) A frame work for extracting conceptual information from document set like definition, explanation, short note, contrast, feature, process, reasoning etc.

b) A frame work for extracting conceptual information from knowledge base such as semantic net, semantic frame and synthesis of the proper answer.

III. Design And Soluition

The present Question Answering System is designed for answering scientific and technical domain questions starting with words "what", "why" and "how". For answering to the question by this system, it is necessary that the question sentence should contain relevant keywords within it.

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 in it. The concepts are the entities about which the question is all about and the attributes define what is needed to be asked/ known about the concepts. For example in the question "why semiconductor have valency 4", the concept "semiconductor"- defines that the question is about semiconductor and the attribute "valency" defines what is needed to be asked or get to

be known about the concept "semiconductor". The concepts in the question are the entities represented as keywords, which are of course the Nouns. The attributes are noun like "resistor", "resistance" and verb like "resistivity". Thus noun, verb found in question is assumed to be an attribute.

A technical or scientific question may contain one or more concepts and attributes within it. The system produces answer with the assumption and philosophy that producing the relation among the required concepts and description of relation between attribute with concept, if no relation is found then mere description of the concept will obviously covers lot of the answer part in it.

The relation among the concepts in a domain can be defined by a conceptual map. The concept map can be made by semantic net. The semantic net defines how one concept relates with another concept. It is a directed graph consisting of vertices/nodes representing the concepts and edges representing the relationship between the concepts. There is no standard set of relations in semantic networks, but the following relations are very common:

- INSTANCE: X is an INSTANCE of Y, when X is a specific example of the general concept Y. Example: Johan is an INSTANCE of Human.
- ISA: X ISA Y, when X is a specialization of the more general concept Y. Example: sparrow ISA bird.
- HAS - PART: X HAS - PART Y, when the concept Y is a part of the concept X. (Or this can be any other property) Example: car HAS seat.

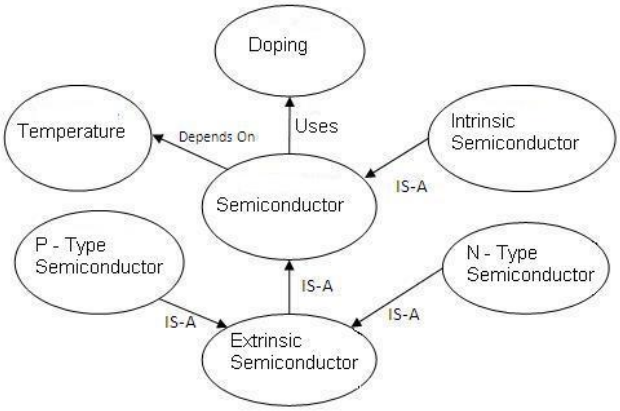

Figure 1. Figure shows a section of semantic net used to represent part of the domain "semiconductor physics".

Every concept in the semantic net has its own set of attributes, which are described by semantic frame. The semantic frame consists of slots, each slot contains fields called facets which specify attributes, attribute's value and description or comment. The description is a brief note/explanation on why the attribute has that value.

To answer a general conceptual question, explanation of its concept/s is necessary. A conceptual question ("why" & "how" type) is of the form – "Why / How ….P……Q ?", where P and Q are concepts, the question can have only one or more concepts. It is found that one concept has an action on other concept or one concept is the cause and the other concept is its result. To explain this kind of conceptual question it is very important to get the relations between the concepts, like P – relation1 – X – relation2 – Y – relation3 – Q, where X, Y are the other concepts. The relationship states that to explain concepts in the question the other interdependent concepts need to be explained. The relationship among the various concepts in a domain is obtained by a conceptual map. For the question having attribute/s along with the concepts like "why / how….P….a…Q ?", the attribute 'a' defines what need to be asked or get to be known about the concept/s. Here the attribute 'a' may be of concept 'P', 'Q' or of any concept present in between the relationship of 'P' and 'Q'. As it is not known, attribute 'a' belongs to which concept, searching the attribute for a match in the attribute set described in semantic frame for each concept present in the relationship of 'P' and 'Q' for a description regarding the attribute and its value can give the answer to the question. Thus for every concept found in question, relation among the concepts are derived, this would explain in what way the concepts are related and thus interdependency is found. The attributes found in the question are searched against the semantic frame of each concept in the relationship, to know about what is to be known about the concepts.

A Technical Questions may consist of:

- One concept with no attribute
- One concept with one or many attribute.
- More than one concept with no attribute.
- More than one concept with one or many attribute.

For questions having one concept, the concept is searched in semantic net and the description of the concept is obtained from the semantic frame, which provides the answer to the question.

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

For the questions having one concept and one attribute, the concept is searched in the semantic net. The corresponding semantic frame is searched and in its each slot, both the attribute and attribute's value fields are matched against the given attribute. The matching slots description is extracted as the answer to the question.

In questions 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 question's attributes is searched for a match in the semantic frame of each concept present in the relationship. For match not found 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 is extracted from tagged text

document. Technical subject/domain document content can be categorized into different sections and parts as introduction, definition, brief, explanation, types, property, features, behavior, function, procedure, components, parts, structure, necessity, application, theorem, principle, formula, example, figure, conclusion etc. Based on these respective tags are made to represent the section like:

1. <introduction>……</introduction>
2. <def>……..….</def>
3. <brief>……....</brief>
4. <explanation>…….</explanation>
5. <types>……...</types>
6. <property>……..…</property>
7. <feature>……</feature>
8. <behavior>……...</behavior>
9. <function>……</function>
10. <procedure>…….</procedure>
11. <components>……..</components>
12. <parts>……</parts>
13. <structure>……...</structure>
14. <necessity>……..</necessity>
15. <application>……..</application>
16. <purpose> ……..</purpose>
17. <theorem>……...</theorem>
18. <principle>…… </principle>
19. <formula>……..….</formula>
20. <example>……..…</example>
21. <figure>…….…..…</figure>
22. <conclusion>……..…</conclusion>

Each tag has attribute – "name", "of" with value set as the keyword name (Concept name). The tag should have at least one attribute in it. The attribute "name" defines what is being tagged and "of" define of whom. The tags can be nested. e.g., <def of: semiconductor >A semiconductor is a substance …… e.g., <example of: semiconductor>: germanium, silicon, carbon</example></def>

While processing a "what" question, the question is identified for what it's asking and the corresponding tag is recognized. The tagged documents are searched with the identified tag name and tag's attribute value set as the concept's name. The matching tagged information is extracted as answer.

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 natural language text from user, assuming the same or semantically related question is not found in the FAQ base – the question text is processed. The text processing includes spelling check, word trimming, base word conversion, words stemming, and removal of words with least significance, extraction of concepts and attributes.

The Answer generation process takes the concepts and attribute as input from the Question processing process. The 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 of the question sentence its type is determined as to whether "what", "why" or "how" type. The "why" and "how" types of questions are considered as reasoning based question, the information for which are obtained from the knowledge base and are combined together to form the answer. For question of "what" type, the system identifies the tag and the tagged information is extracted from the document set else if no tag is identified the answer is obtained from knowledge base.

The Answer authentication process obtains the unauthenticated answer from Answer generation process and sends it to the expert for authentication. The expert authenticates or corrects the answer with minor alterations. The process saves the question and its authenticated answer into the FAQ base and also sent the answer to the corresponding user, who has raised the question.

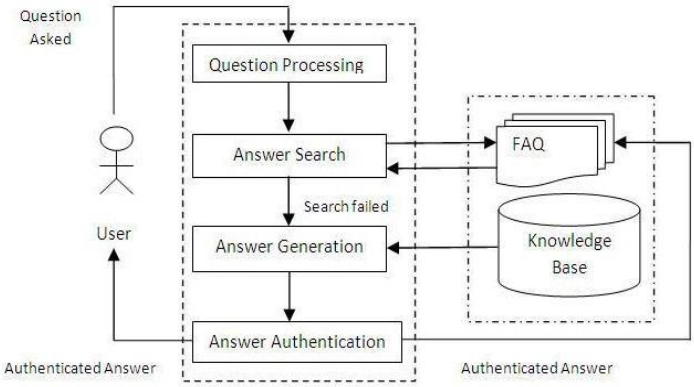

Figure 2. Framework for automatic answer generation system for FAQ based QAS.

IV. ALOGORITHM

Step1 The question is taken as input in the form of string.

Step2 The question is segmented into individual words using white space as delimiter.

Step3 Each segmented word is matched in the domain vocabulary. If the word does not finds a match in the lexicon, its set as wrong spelling.

Step4 If each word spelling is found correct then go to step 8.

Step5 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.

Step6 words with its wrong spelling and the correct spellings are displayed as possible solution of the word to the user for input.

Step7 Correct spelling words are taken as input from the user as a replacement for the wrong spelling words

Step8 Each segmented word is matched in the domain vocabulary, for a match found it is converted to its base word/root word.

Step9 The words that are found in the domain vocabulary are identified and tagged as "KEYWORD".

Step10 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]

Step11 A normalized question is obtained by eliminating pronoun, auxiliary modal verb, adverb, adjective, article from the question and selecting only keywords, noun and verb.

Step12 If the first word is "what". Set the question type to "what_type" else go to step 17.

Step13 The normalized question is searched for tag words.

Step14 If the tag word is not present, go to step 18.

Step15 If tag word is present, the tag word is searched in the tagged text file/document against the keywords present in the normalized question.

Step16 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 ".

Step17 If the first word is "why", "how", set the question type to "why_type".

Step18 Each word of the normalized question is matched with every nodes of the semantic net.

Step19 If found equal, the name is added to the nodes array else the name is added to the property array.

Step20 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).

Step21 The duplicate information in the information array is removed.

Step22 The information stored in the information array is displayed.

a) Algorithm for search (nodes array, property array)

Step1 for node array length = 1 and property array length = 0, call search1(nodes array[0]).

Step2 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 ,

a. Display "Since property array found no match in the properties of nodes array, the description of will probably give the result"

Step3 call search1(nodes array[0]).

Step4 for node array length >1 and property array length >= 0 Call search3(nodes array, property array).

Step5 For node array length = 0 , display "The keywords / words not found in the knowledge base".

b) Algorithm for search1 (element): searching single node in the semantic net

Step1 The element is searched against every node in the semantic net.

Step6 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.

c) Algorithm for search2 (element, property): searching single concept with single property in the semantic net

Step1 The element is searched against every node in the semantic net.

Step2 If the element searched is found, the property is matched for each attribute in the corresponding semantic frame.

Step3 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

Step4 If base node found, go to step 2, else set flag to false.

d) Algorithm for search3 (nodes array, property array) : searching multiple nodes with multiple property in the semantic net

Step1 The elements are searched against every node in the semantic net.

Step2 The routes between the nodes are mapped.

Step3 If the route do not exists, set flag to false.

Step4 if flag = true, Relation between the adjacent nodes in the route is stored in an information array.

Step5 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 arraycall 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])}

Step6 If flag = false, display "Route between the nodes doesn't exist"

V. Evaluation And Conclusion

An initial evaluation is performed on a technical domain with a set of questions. The question are made strictly using "what", "why" and "how" as the initial word of the sentence.

After analyzing the answer generated it is found that the answer contains extra information in it. Based on the relevance of information contained in the answer, the answers are categorized into 2 categories as below:

Category "A": The answer contains exact pieces of information,

Category "B": The answer contains exact piece of information with addition to that one or more pieces of extra information.

The system is evaluated with 50 questions each on "what", "why" and "how" types, the following results are obtained.

TABLE I. NNUMBER OF QUESTIONS OF WHAT, WHY AND HOW TYPE IN CATEGORY A AND CATEGORY B

	A	B
What	41	9
Why	35	15
How	30	20

The initial evaluation result shows the feasibility of building QAS with this framework, which can automatically produce answers to conceptual questions in a science and technical domain. Thus can be used as tool to answer the Frequently Asked Question and helps in building the FAQ base.

Use of knowledge base (semantic net and semantic frame) and tagged document set is found very helpful. The use of tags has solved the problem of extracting answers from text to an extent as it annotates the text, sentences or passage with meaningful information and thus helps in extracting the required piece of information correctly.

For understanding the meaning of question use of concept present in question is quite effective as it requires much less Natural Language Processing tasks. The use of knowledge/conceptual map is definite requirement as this technique is the very effective in understanding the various concepts present in the question and their relations. 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 generates message for every instance when

a) The system does not found the required tag for a concept in the document.

b) Searched concepts are not found in the concept map [semantic net].

c) Searched concepts are not related in the concept map [semantic net]

d) Attribute in question don't match against the attributes or its value in the semantic frame of the required concept.

The messages produced by system helps the expert in changing and upgrading the knowledge base [semantic net and semantic frame] and adding tags and other required information to the documents so that the system can suitable produce the right full answer for the future queries.

The tags used are on a single technical subject and needs to be tested more on other domain, so that tag set can be made complete and a standard on tag can be obtained.

REFERENCES

[1] Chun-Chia, W., Jascon, C. H., Che-Yu, Y., Husan – Pu Chang, "A Question Answering System Approach for Collaborative Learning", 10th International Conference on Computer Supported Cooperative Work in Design, IEEE Conferences, May, 2006, pp. 1 – 5.

[2] Boldrini, E. ,Ferrández, S., Izquierdo, R., Tomas, D., Ferrandez O., and Vicedo J.L, "A proposal of Expected Answer Type and Named Entity annotation in a Question Answering context", 2nd Conference on Human System Interaction, IEEE Conferences, May, 2009, pp. 318-322.

[3] Keliang, J., Xiuling, P., Zhinuo, L.,"Question Answering System in Network Education based on FAQ", The 9th International Conference for Young Computer Scientists, IEEE Conferences, Nov, 2008, pp. 2577- 2581.

[4] Parik, J., Narashima Murty, M., "Adapting Question Answering Techniques to the Web", Proceeding of the Language Engineering Conference, IEEE Conferences, Dec, 2002, pp. 163-171.

[5] Qinglin G., "Question Answering System Based on Ontology" , 7th World Congress on Intelligent Control and Automation, IEEE Conferences, June,2008, pp. 3347-3352.

[6] Yongping, D., Ming, H., "Pattern Optimization and the Application in Question Answering", Eighth International Conference on Intelligent Systems Design and Applications, IEEE Conferences, 2008, pp. 35 – 39.

[7] Zheng-Tao, Y., Yan-Xia, Q., Jin-Hui, D., Lu-Han, Cun-Li, M., Xiang-Yan, M., "Research on Chinese FAQ Question Answering System in restricted domain", International Conference on Machine Learning and Cybernetics, IEEE Conferences, Volume 7, Aug, 2007, pp. 3927 – 3932.

[8] Xiaobo, W., Wei,C., Weicun, Z., "Design and Realization of Intelligent Question Answering System Based on Ontology", International Conference on Control, Automation and Systems, IEEE Conference, Oct, 2008, pp. 1313 – 1316.

AUTHORS PROFILE

Saurabh Pal is working as faculty in the Dept of Computer Science and Information Technology at Bengal Institute of Technology, Kolkata, India. He received his B. Sc. from University of Calcutta in 1998, M.Sc. IT from Allahabad Agricultural Institute – Deemed University in 2006, and M.Tech – IT [Courseware Engg] from Jadavpur University in 2010.

Sudipta Bhattacharya is working as Asst Prof in the Dept of Computer Science and Information Technology at Bengal Institute of Technology, Kolkata, India. He received his B.Tech from West Bengal University of Technology in 2006, M.Tech – IT from Bengal Engineering and Science University in 2009.

Indrani Datta received her B.Sc Software System from North Bengal University in 2004, Master of Computer Application from West Bengal University of Technology in 2008. Her primary area of interest is Natural Language Processing, Information Extraction.

Arindam Chakravorty is presently working as Asst Prof in the Dept of Information Technology at St Thomas' College of Engineering and Technology , Kolkata, India. He received his B.Tech(Electrical), M.Tech IT(Courseware Engg) from Jadavpur University.