# Used simple rule based approaches english language essay

Linguistics, English

Working in medical domain avoids many ambiguity problems and, but even in a scope some problems can occur. Information requirement of health professionals have been researched for more than two decades. Early surveys (Strasser 1978) required information on new developments in the field of specialties and government regulations relating to health care [47]. Mostly the sources of information were journal papers, colleagues, and books. Further observation studies concentrated on the types of information needs the clinicians have and the type of questions arising while attending the patients. In a review of the research of information needs of health professionals'. Smith (1996) identified six categories of information needs, and remarks on the nature and importance of the required information [48]. Out of these, Medical Knowledge is the center of our thesis. Richardson and Wilkinson made a distinction between the background and the foreground medical knowledge [49]. The background knowledge is a general knowledge about the basic concepts of a disease. Information needs related to the background knowledge will generate information requests in the form of wh-type questions (who, what, where, when, why, and how) such as What is this trouble?, What causes it?, What are the signs and symptoms?, What treatment choices that exist? These questions are better answered from textbooks and regularly updated systematic reviews. Such resources collected and compiled by specialists are often called secondary sources, or secondary literature unlike the primary sources that report original results of the clinical research. More often, practicing clinicians miss the foreground knowledge: knowledge about the choice of therapeutic interventions, the best diagnostic check-up for a disease, or the best intervention strategy for a

particular patient. These questions can be answered by systematic reviews if they are updated regularly, and if the questions are common. In case of non-availability of secondary resources, the online databases which provide results of clinical trials and observations are used as a source for answers. ◦ Are the answers substantial enough and provide the best ways to answer the foreground questions?◦ Do the existing resources contain enough information capable of answering these questions?◦ To what extent is the available knowledge accessible and used by clinicians? The first question was answered positively. The second question was surveyed in several experiments in which physicians or medical librarians were given a set of foreground clinical questions and were asked to find answers using medical literature and online databases. Gorman et al. (1994) found medical librarians deemed 88% of the clinical questions appropriate for MEDLINE, and found information judged relevant by physicians for 56% of the questions [50]. Clear answers were got for 46% of those questions. Clinicians evaluated 40% of the answers as having impact on their patients, and 51% of the answers as having impact on themselves and their practice. Similarly Giuse et al. (1994) answered 87% of clinical questions generated from patients' charts using online bibliographic databases [51]. In a study, two experienced physicians found answers to 75% of the questions. Koonce et al. (2004) found secondary sources capable of answering 20% of the foreground questions and 47. 5% of the background questions, which signifies the importance of the primary medical literature [52]. We know that online databases can answer at least half of the clinical questions, researchers analyzed at what extent clinicians use these resources. De Groote and

Dorsch (2003) surveyed medical students, residents and faculty and found that 53% of the users searched MEDLINE at least once a week, 72% of the surveyed users used the online resources for patient care [53]. The key findings of the national audit of CQA Services conducted by Doctors. net. uk (Bryant 2005) indicate that 78% of the doctors consulted colleagues as their first source in answering clinical questions [54]. Gorman and Helfand (1995) found that doctors pursued less than a third of the questions and found answers for less than a quarter [55]. Problems in answering clinical questions were studied and found that physicians quested for answers only to 55% of the questions due to failing to recognize an information need, uncertainty regarding the existence of an answer, preference for convenience rather than appropriate resources, and lack of skills formulating questions and search strategies. The latter issues are addressed in the practice of Evidence Based Medicine. Evidence-based medicine (EBM) approaches to bridge the gap between the care that a patient will get and the best possible care in a given situation as determined by systematic research. EBM provides guidelines for developing clinical questions and translating these questions into successful information requests. For each of these clinical tasks, there are four elements of a question: PICO plays a role in the question and search formulation. Detailed descriptions of the three components of the EBM-based semantic domain model are. Physicians’ everyday activities require information related to common health problems and adequate diagnostic, therapeutic and preventive services. Sackett et al. (2000) find eight task categories that described these activities: clinical findings, diagnostic tests, etiology, differential diagnosis, cause, treatment, prevention, and self-

improvement [56]. The Task Force ratingIt is found that the distinction between the differential diagnosis and etiology questions is not always obvious. For example, in a scenario based assessment of physicians' requirements, What is anemia? Was found as a question about diagnosis as the patient's test results showed abnormal clues indicating anemia, but What is the reason of gastritis? Was identified as an etiology question because gastritis was absent in the test results (Seol et al. 2004) [57]. The task of improvement is one of the aims of the evidence based medicine practitioners. It is absent in the AHRQ rating scheme, as any questions of improvement got during the practice of medicine could fall into one of the above four categories. This task pertains to improvements in clinical practice through improving their learning skills, as shown in the following question: To improve the understanding of the pathophysiology of ascites would I get more from spending an hour in the library reading a textbook or spending 25 minutes on the ward computer looking at the same textbook? (Russell) [58]. An answer to this question depends on various factors outside of a clinical scenario. To improve the act of formalizing information needs of clinicians, Richardson et al. (1995) proposed PICO for constructing well-formulated questions [59]. These questions determine the patient and/or problem, a planned intervention (e. g., a treatment or a diagnostic test), a possible outcome of the intervention with a comparison intervention as follows: Patient (population)/Problem: identify information about the patient or a group of patients and the problem that requires clinicians' care. Such information is routinely got during collection of preliminary case history and complaints given by patient and the subsequent diagnostic work-up.

Intervention: it is the procedure, agent or other clinician’s act of interfering with a case to modify it or with a process to change its course that is being monitored to either a single patient or a group of patients. Exposure: it is an alternative to the intervention slot of the PICO frame developed to include the etiology (harm) questions. The exposure also reflects the phenomenon of interfering with a patient’s condition, but the actions are that of harmful agents, e. g., excessive exposure to cigarette smoking. Comparison: It provides a frame of reference for an intervention, for example, in an alternative intervention, another method of administration or pattern of dosage or a different timescale. Patient outcomes of a planned intervention could be measured against a comparison. Outcome(s): it summarize the effect of an intervention or an exposure to patient or population, focusing on patient oriented results such as few side effects, increased survival rates, restoration of functions, etc. Questions containing these components are deemed to be more " answerable." Although the question construction process was initially thought to answer questions arising with respect to the therapy, it was later adapted for all clinical tasks in a series of articles (JAMA) (Guyatt, Sackett, & Cook 1994) [60]. Each article aims on constructing PICO representations of clinical questions for a major clinical task. Scenario: A 75 year-old woman with controlled hypertension and a history of non-valvular atrial fibrillation resistant to cardioversion wants to get whether the benefits of long-term anticoagulants (to reduce the risk of embolic stroke) outweigh her risks (of hemorrhage from anticoagulant therapy). The authors selected non-valvular atrial fibrillation from the overall description of the patient and her problem to populate the Problem slot of the frame. Based on the

knowledge, the Intervention slot was populated with warfarin. Trustworthy evaluations of drug effectiveness formed basis of comparisons with other drugs likely to be effective, or no treatment that often amounts to treatment. In this case authors do populating the Comparison slot with placebo. Finally, there are several Outcomes of interest: risk involved of emboli (including embolic stroke) and the risk of the complications of anticoagulation. Although the JAMA article (Jaeschke, Guyatt, & Sackett 1994) discusses about PubMed search strategy and identifies significant concepts that need to be included, there are no direct recommendations found for assigning concepts to the PICO frame slots [61]. A recent article offers the following example: Question: accuracy of an increased respiratory rate to detect pneumonia in most children's presenting to a clinic with respiratory symptomsP: In children's with upper respiratory symptomsI: is determining the respiratory rateC: as effective as the chest x-rayO: in detecting the pneumonia? However in an EBM training course conducted by the British Medical Academy (BMA) the suspected problem fills the problem slot and the outcome slot remains as empty: Q: What are the diagnostic tools for the screening of prostate cancer in young males, and how effective are they? Population /Patient /Problem: Prostate cancerIntervention: ScreeningOther limits: Male Adolescent AdultThis frame instantiation not only present the difficulties in finding an appropriate slot for the hypothesized disease, it also suggested the need in separating the Patient and the Problem descriptions by getting patient's gender and age into a separate field. The JAMA article on etiology questions (Levine et al. 1994) express an asthmatic patient asking his doctor about risk of death associated with beta-adrenergic agonists in the

cure of asthma, but does not provide a filled PICO frame [62]. The BMA tutorial gives the following example: Question: What is the risk of psychiatric illness upon taking the antimalarial drug Mefloquine? Population /Patient /Problem: Psychiatric illnessIntervention: MefloquineAccording to the example, in the original JAMA etiology question, asthma could be assigned to the population/problem slot frame; beta-adrenergic agonists to the intervention; and adverse effects to the outcome slot of the frame. There is however a difference in allotting an existing disease - asthma to the problem slots, and a potential result of exposure to the drug to the same slot, as in Psychiatric illness. In the asthma example, beta-adrenergic agonists are an intervention with respect to the disease, and an exposure with respect to the patient outcome, this case could have risk of death. What if the patient condition is an onset of a new pathologic condition, for example psychiatric illness? Such EBM sources as The Pediatric Residency Curriculum Handbook, University of Illinois at Chicago, advice placing a potential disease into the outcome slot: Patient/Problem: Controlling for various confounding factors, do otherwise healthy childrenIntervention/Exposure: exposed in utero to the cocaineComparison: compared to children not exposed, Outcome: have an increased chance of learning disabilities at age six years? Prognosis refers to the possible outcomes of a disease and the frequency with which they occur. In the JAMA example (Laupacis et al. 1994), the question for prognosis is a man’s inquiry about an Alzheimer’s patient prognosis, and whether he is likely to die soon [63]. In this case, Alzheimer’s disease considers the problem slot, and death - the outcome slot. There are controversial recommendations in getting PICO frames for prognosis questions as well.

The Chicago handbook gives the following example: Patient/Problem: In children having Down’s syndrome, Intervention: Is IQ main prognostic factorOutcome: in predicting Alzheimer’s later in life? The BMA tutorial suggested placing the motor neuron disease into the problem slot for the following question: What is the short / long term prognosis for a young adult diagnosed with motor neuron disease? Once the question is processed, the JAMA series of articles advices searching PubMed using all terms in the PICO frame. The articles recommend augmenting the search with a definition of the clinical task, for example, prognosis for prognosis questions, and cause for etiology questions. In addition to the question and search formulation strategies, JAMA articles provide a framework for evaluation of relevance and quality of the articles. The major points in evaluating articles are: whether the results are valid, and whether the information is related to the patient’s condition that led to the question. The second point is verified through the PICO structure. The closer the problem, patients, interventions, and outcomes described in the article are related to the clinical scenario, the more might be its significance to the patients. Each component of the EBM-based model has been used in information retrieval separately and to a different extent. One of the successful applications of one of the three basic components is a clinical-task specific query expansion. This research is actively done by the Hedges Project (Wilczynski, McKibbon, & Haynes 2001) [64]. The accuracy of the manually defined search strategies for therapy, diagnosis, review, prognosis, causation (etiology), economics, cost, and clinical prediction guides was done on 49, 028 articles indexed for MEDLINE. The best performing terms were selected from 4, 862 unique terms. These

terms are present for query expansion in the form of Clinical Queries in PubMed, a service of the U. S. National Library of Medicine that gives access to over 16 million citations from MEDLINE Database. is a " clinical study" as judged by a categorizer having trained on publication types and other features of 7000 MEDLINE citations. Information retrieval research in biomedicine parallels with the research on information retrieval in general. Some of the initial studies were conducted at the National Library of Medicine in anticipation of computer-based retrieval systems. For example, Winifred Sewell (1964) surveyed library’s controlled vocabulary used for indexing in anticipation of the Medical Literature Analysis and Retrieval System (MEDLARS) [65]. Online access to a subset of references present in the MEDLARS database became available in 1971 in the form of MEDLINE (MEDLARS Online). Sewell regarded the controlled vocabulary terms, called Medical Subject Headings, " as directional signals which, with other headings, served to locate the essence of a particular paper or book in the universe of medical information." She found that through greater coverage and deeper indexing a computerized system would increase the requirement for specificity in descriptors and delineation of hierarchical relationships useful for search purposes. Sewell identified four broad groups that have the majority of the vocabulary terms: Anatomical Terms; Organisms; Diseases; and Chemicals and Drugs. The desired level of specificity for adding concepts to the top level hierarchies was done by frequency of appearance of concepts in journal articles and by the ability of a specific term to retrieve about a 100 citations from the 1961 collection of articles. By the theory of sublanguages proposed by Harris, such computation oriented corpus-based

definition of the controlled vocabulary is made possible due to the structure and regularity of technical languages (Friedman, Kra, & Rzhetsky 2002) [66]. This notion of the special biomedical sublanguage determined several directions in domain-specific information retrieval explored in addition to the mainstream research techniques confined to specialized literature. Domain specific research is largely affected by existing resources such as databases of genomics information, For example: Gene Ontology, and MEDLINE. Biomedical journal articles are different from documents comprising the widely used Information Retrieval (IR) collections (such as TREC collections that primarily include news stories) not only in the specifics of the biomedical sublanguage but also in their structure.. Purcell et al. (1997) uses the complex structure of medical articles in three hierarchical context models (clinical research articles, case reports, and reviews) for medical document representation, and find a number of elements, which the authors refer contexts, that characterize each of the structures [67]. Purcell et al. proposed annotating each sentence in an article with some context, for example, experimental findings in the results of a clinical research article, for the purposes of context-based information retrieval, and got significant improvement in precision of full-text searching at fixed levels of recall. Domain knowledge resources are widely used in the production information retrieval systems (PubMed, Ovid), in natural language processing (NLP), and in information retrieval research. The following sections give some of the key domain knowledge-based techniques. In addition to query refinement for clinical search developed by the Hedges project, many researchers studied query refinement techniques that are useful in ad hoc retrieval. Srinivasan

(1996) compared three sources for having blind relevance feedback (query expansion using additional terms from the initial set of retrieved documents [68] In Srinivasan’s study, initial search was compounded using only controlled vocabulary terms, only free text terms, and a combination of the two sources. Improvements in average precision at 11 standard recall points in the range of 9% to 119% were got for individual queries, with an overall improvement of 16%, primarily due to the controlled vocabulary feedback. The significance of MeSH terms and the slowness and cost of the current manual indexing process led to a number of studies of automatic extraction of the UMLS concepts from medical text. Other methods include mapping of query terms onto MeSH terms through a common semantic representation based on 3400 simple atomic concepts such as " heart" (Zieman & Bleich 1997), restricting UMLS concept matching onto noun phrases, or first generating all possible UMLS concepts for each of the text tokens and then using syntactic and semantic filters to eliminate irrelevant candidates [69]. A related effort – protein and gene name identification – got into much attention recently, following the growing interest in the " omics". " Omics" include genomics study of a living organism in terms of its sequence of its genome; proteomics – study that focuses on determination of physiological roles of proteins and their structure; and a relatively new field metabolomics – study of metabolites that show the end product of gene expression. Entity identification methods for " omics" comprise of dictionary look-up, rule-based term recognition, machine learning, and hybrid approaches. A comprehensive review of gene and protein name recognition techniques is given in (Krauthammer & Nenadic 2004) [70]. One of the first specialized

retrieval systems that demonstrated automatic concept-based indexing and extraction of the UMLS concepts from users' requests was SAPHIRE (Hersh & Greenes 1990) [71]. SAPHIRE uses the UMLS Metathesaurus by fragmenting free text into individual tokens and constructing a list of Metathesaurus terms for each token. The terms were weighted based on their length, overlap with the original text, and the proximity of the original tokens with each other. When compared with regular MEDLINE searches performed by physicians, SAPHIRE performed well for physicians, but was outplayed by librarians using MEDLINE. Chen et al. (2003) augment noun phrase indexing with automatic thesaurus generation in the HelpfulMed system, a Web portal that gives information retrieved from reliable medical domain sources with minimal manual effort [72]. HelpfulMed also gives several presentation modes of retrieval results: in a traditional ranked list, in a self organizing map, and in a list of automatically derived concepts, MeSH terms, and authors, which provides a user an opportunity to search phrases generated from the text, related medical subjects headings, authors, or any combination of the three, thus accommodating users with different information needs and tasks. Question answering encompasses field psychology, philosophy, linguistics, education, computer and library science. As a consequence, studies of the artificial intelligence, in particular natural language processing, and information retrieval aspects of question answering get benefit from knowledge acquired in other disciplines. Philosophy and psychology provides flavour of the question answering process. According to Singer's (2003) review of the theories leading to understanding of the process, its first stage is the encoding of the question

meaning. Singer follows Kintsch’s tradition in presenting questions as propositions [73]. He also points out that successful question comprehension and answering depend on understanding of which parts of the question have information known to the person asking the question, and which part is the request for desired information, i. e., the focal idea of the question. Identification of the question focus is related to the listener’s knowledge. Lehnert (1977) gives the importance of finding focus in an implementation of a prototype question answering system SAM [74]. SAM, which tried approximating human cognitive process, answered questions about stories depicting eating in a restaurant. It used a sentence analyzer, a script application mechanism to build a memory for the story, and procedures for locating answers to questions in its memory. This prototype presents the first generation AI question answering programs. The five classes of systems identified in this taxonomy are strongly related with the types of questions and the available test collections. Class 1 system which is capable of processing factual questions and typically extract answersusing keyword matching. Class 2 systems which use semantic alternations, world knowledge axioms and simple reasoning to relate snippets of text containing answers with the questions. Class 3 systems which generate answers to list, script, or template-like questions from parts found in several documents. Class 4 which interactive systems answer questions in the context of previous interactions, which involve complex reference resolution. Class 5 systems which answer speculative questions, which involve knowledge extraction from relevant documents and case-based, temporal, spatial and evidential reasoning. Since for the most evaluations done at the Text

Retrieval Conferences (TREC) (Voorhees & Tice 1999; Voorhees 2003), NTCIR and CLEF focused on the fact-based questions, many Class 1 systems that successfully utilized surface text pattern matching have been developed (Brill et al. 2001) [76]. The major advantage of their approach is simplicity: the use of surface patterns needs minimal processing, resources, and knowledge engineering compared to the other types of systems. Closed domain question answering has recently gained the interest of researchers. The definition of the closed domain varies from working in a specific domain to using closed document collection restricted in size and subject. The term restricted and closed-domain are used interchangeably, but Benamara (2004) defined it to be broader in terms of subject coverage, for e. g., tourism covers various subject areas, e. g. accommodation, transportation, etc., as opposed to Unix manuals. Interestingly, there seemed no reported attempts to apply systems developed for open domain to the closed domain by researchers also with successful open-domain systems [77]. However, closed domain systems have started explorations in other domains, for example, the ExtrAns system developed to answer questions using Unix and Aircraft Maintenance manuals is re-targeted to answer genomics questions (Rinaldi et al. 2004)[78]. An interesting validation of the PICO framework originates from a study of the informal consultations between 60 primary care physicians and 30 specialty physicians. In this study, e-mailed questions were remained unanswered if they identified a proposed intervention and a desired outcome. The comparison had no effect on the answer (Bergus et al. 2000) [79]. Booth and O’Rourke (2000) developed application of the PICO framework for retrieval of documents and found that structuring abstracts

according to PICO improved the accuracy of search for clinical questions compared to unstructured single paragraph abstracts [80]. Niu and Hirst (2004) applies PICO framework to search a database of reviews that summarize and appraise clinical evidence, and reports preliminary results of outcome identification in these reviews [81]. However it is unlikely that findings from a survey of peer-reviewed compilation for 200 medical conditions generated by a limited number of specialists will scale to MEDLINE abstracts. For example, a very specialized source allows using terms like comparison and dependency as indicators of patient outcomes. But, the term comparison can be found in 417, 589 MEDLINE abstracts, only in the title, i. e. Comparison of preoperative anxiety in reconstructive and cosmetic surgery in patients. In general database, these terms could lose their predictive power suggesting that simple cue words are only the beginnings of a solution. The third component of the EBM-based model, the strength of evidence, is present in MEDLINE users since 1991 in the form of manually assigned Publication Type controlled vocabulary terms, e. g., search can be restricted to the strongest example in the form of meta-analysis and randomized clinical trials. Several surveys exploit this component of the EBM-based model in document ranking. This component is taken into consideration in summarization (Fiszman, Rindflesch, & Kilicoglu 2004) that retrieves abstracts with the most effective publication types as the first step of the process [82]. Similarly, McKeown et al. (2003) personalize find results to a patient profile taking into account whether a document is a " clinical study" as judged by a categorizer trained on publication types and other features of 7000 available MEDLINE citations [83]. Question Answering

System (QAS) performs the task of, given a user query expressed in Natural Language (NL), retrieving its correct answer in the form of compact text from large text based documents. The objective of this task is to reduce the amount of time required by the user in seeking information in which the system processes and analyzes information in documents and returns the piece of relevant information to the user. There is a vision that in the future machines will have an interactive communication with humans, by answering a wide range of questions. Typically, a traditional QA system consists of three main steps processed in a sequential manner, namely question analysis, search, and answer selection (Hovy et al., 2000) [84]. A QA system returns an accurate answer to a question, if it can determine the type of the question, such as factoid, list, or ‘ other’. This step is very important for the next steps, because answers are extracted according to the question type. In the search step, a question is redeveloped according to its question type and target, and then answer or a document corpus is searched to get all matched answers. Finally, an answer is selected and returned to the user. There are two strategies for answer retrieval: a table look-up strategy and an IR-based strategy. The first strategy is based on the observation that answers for some question types can be expected, and hence, the corpus can be analyzed off-line. The second strategy extracts answers from relevant paragraphs, that contain keywords from the question. Tables in an open-domain QA system contains relations between named entities, such as names of persons, organizations, dates, locations, currencies and amounts. For example, a state relation links a city name with a country name; a birth date relation links a person name and a date, and so on. The extraction of

these relation types are the most important tasks in the table look-up method. On the other hand, IR techniques are successful for retrieving relevant documents on the Web. However, the overall time to process questions, to submit and retrieve answer from a Web search, and then to extract relevant answers is relatively long. It would be more efficient if answers to frequent and simple questions, such as to factoid questions, could be looked up from the precompiled tables. This section deals with the Design and Implementation aspects of the suggested solution. Here Java, a programming language tool is used as it is truly Object-Oriented Programming language. The Proposed solution should be easily extendable and modifiable. A High Level Design is shown belowThe interface is the platform where a user interacts with a map and submits his question and receives answer from system. The user interface is significant element of a QAS. Question Answering System is the core of overall system which interacts with different components user interface, Word Net, Parser, Name entity Tagger and generates output as answer to the user’s query. Question Answering System (QAS) performs the task of, given a user query expressed in Natural Language (NL), retrieving its correct answer in the form of compact text from large text based documents [4], [6], and [16]. Our question answering system does extract most relevant answer to the question from the large document which contains information related to medical domain. Our approach of question answering system firstly classifies the query, then extracts the candidate answers and finally ranks these candidate answers. When the user posed the question, the system extracts the answer information from the tagged document passages in terms of

named entities. From these named entities some will represent the candidate answers of the question. The other part of our system finds out the semantic relation between question and candidate answers and assigns the weights to the candidate answers and ranks the candidate answers according to the weighs which gives the best accuracy with the use of Word Net [17] before generating final answer to user query. The parser is component of compiler or interpreter which act as the natural language processing tool that does syntactic analysis thus considers the grammatical construction of sentence, it can find out which words are used as subject, verb or object , phrases [6], [37]. We used Stanford Parser, a tool which generates the Parts of Speech (POS) of each word of the inputted user query and the candidate answers selected from documents. Word Net is generally used in many QAS as it happens to be a very helpful tool when it deals with words. The Word Net characteristics are utilized to point out the relationship existing in between words of user query and data source. Word Net is the massive lexical database consisting of English words. We used OAK, a named entity tagger which tags about 150 named entities. These entities are arranged in hierarchyHere we go through system overview which consists of external tools that are being used in our system. In our question answering system, we are using a Semantic Approach for extracting the answer from the text documents. In this method sentence are selected on the basis of concept rather of specific words. The document tagging module of our system tags important information from the passages generated by the information retrieval system. Our system uses information extraction techniques to extract important information to help in question answering. In

this chapter there is information on the following ways for tagging information in documents:• Tokenization and splitting of sentence• Part of speech tagging• Chunking of part of speech• Word sense tagging• Word dependency tagging• Named entity tagging of document• Co reference resolution taggingThese tags deal with different divisions of processing. The first is tokenization which separates the characters into words and punctuations. The rest are part of speech tagging, word sense tagging, and named entity tagging. These tags give meaning to each word. The next higher tags are chunking of part of speech and co reference resolution tagging. These tags make group together with words that relate to each other. The highest tag which we are considering is word dependency tagging that tags the relationship between words. Following will be an overview of the systems our program uses to tag the documents. Then, each of the following sections will define systems that can tag a document. WordNet (Fellbaum, 1998) is not a system that tags documents but we use its features to help tagging of documents and with other modules in our system. WordNet and its features could be referred to frequently in the following chapters. WordNet is a lexical resource with information about words and their relationships. Words are grouped in one of these four categories: nouns, verbs, adjectives and adverbs. WordNet is utilized in most question answering systems as it is a useful tool when dealing with words. For each word, WordNet stores each sense that the word is used in. Each sense of a word also refers to a synonym set (if that sense of the word has one or more synonyms). e. g., the noun auto has five senses and the synset and WordNet definition for each of them are: car, auto, automobile, machine, motorcar 4-

wheeled motor vehicle;; " he needs an auto to get to work" car, railcar, railway car, railroad car a wheeled vehicle adapted to the rails ofrailroad; " four cars had jumped the rails" cable car, car a conveyance for users or freight on a cable railway; " user’s tooka cable car to the top of the mountain" WordNet also includes hyponym and hypernym of words. A hyponym for a word is a set of objects that are instances of that word. For example, the hyponym set for car will include the many different types of cars such as:• ambulance• station wagon• bus• cab• convertible• coupe• cruiser• hardtop• hotrod• jeep• limousineWords found within the hyponym set of a word could also have a hyponym set, for example berlin is a type of limousine with a glass partition in between the front and back seats and will be in the hyponym set of limousine. A hypernym set is just opposite of a hyponym set. Sense one of car has a hypernym set as motor vehicle, which is a kind of wheeled vehicle, which is an entity. These sets have relationships in between two words and are significant in question answering. With hypernym sets and hyponym sets a hierarchy could be formed, with hyponym set being more refined and hypernym being less refined. OAK System (Sekine, 2002) was build at New York University and can tag documents in many ways. The OAK System has the ability to tag documents in many ways:• Sentence Splitter• Tokenizer• Part of Speech• Chunker• Named EntityBefore text is tokenized and split into sentences, it is just a string of characters. Tokenization is splitting a string of characters into lexical elements such as words and punctuation (Jurafsky and Martin, 2000, page 298). Sentence splitting separates the words and punctuation into their separate sentences (Jurafsky and Martin, 2000, page 180). This involves a

system probabilistically determining if certain punctuation, that can be used to finish a sentence, is in fact used to end the particular sentence. For example, a period can be used in an acronym and can also be used to end a sentence, or to be more complicated a sentence can end with an acronym with the last period performing both. This is the first step before further processing can be done to the documents. Oak System use these techniques before further tagging documents. Each word in a sentence is divided as a Part Of Speech (POS) that relies on the way the word is being used. e. g.,, the word fax could be used as a noun (Did you receive that fax I sent you?) or as a verb (Could you fax him that report?)Manually tagging a collection of documents, or even a single document, with these tags is very time consuming. There are most systems available that can tag documents with their part of speech effectively. To be consistent, systems use universal tags for different parts of speech. We are using different tags of the Penn Treebank POS tag set (Marcus, Santorini, and Marcinkiewicz, 1994) as Treebank was used to train the OAK system. e. g., consider a sentence such as: Tag Description ExampleCC Coordinating conjunction and, but, orCD Cardinal number 1, 2, two, 44DT Determiner a, the, anEX Existential there thereFW Foreign word moi, coupe, carpeIN Preposition/subord. conjunction in, on, byJ J Adjective red, mean, goodJJR Adjective, comparative faster, closer, tallerJJS Adjective, superlative fastest, closestLS List item maker 3, 1, TwoMD Modal should, can, mayNN Noun, singular or mass frog, dog, lampNNS Noun, plural frogs, dogs, lampsNNP Proper noun, singular CNN, MaryNNPS Proper noun, plural CarolinasPDT Predeterminer all, bothPOS Possessive ending 'sPRP Personal pronoun /, she, youPP$ Possessive pronoun

their, yourRB Adverb slowly, neverRBR Adverb, comparative slowerRBS Adverb, superlative slowestRP Particle up, offS YM Symbol (mathematical or scientific) +, %TO to toUH Interjection um, ah, oopsVB Verb, base form sitVBD Verb, past tense satVBG Verb, gerund/present participle sittingVBN Verb, past participle satVBP Verb, non-3rd ps. sing, present sitVPZ Verb, 3rd ps. sing, present sitsWDT w/z-determiner which, thatWP w/z-pronoun what, whoWP$ Possessive w/z-pronoun whoseWRB w/i-adverb how, whereTag Description Example# Pound sign #$ Dollar sign $Sentence-final punctuation .?! Comma j; Colon, semi-colon > * ' • •(Left bracket character) Right bracket characterStraight double quoteLeft open single quote iLeft open double quote iiRight open single quote jRight close double quote )≫e. g., following sentence is parsedA fractal is a pattern that is irregular, but self-similar at all size scales; e. g.,, a small patch of ground may have the same general appearancebecause a larger patch or even a huge area seen from high above. This sentence with POS tags would be: A/DT fractal/NN is/VBZ a/DT pattern/NN that/WDT is/VBZ irregular/JJ ,/, but/CC self-similar/JJ at/IN all/DT size/NN scales/NNS ;/:/IN e. g.,/NN ,/, a/DT small/JJ patch/NN of/IN ground/NN may/MDhave/VB the/DT same/JJ general/JJ appearance/NN as/IN a/DT larger/JJRpatch/NN or/CC even/RB a/DT huge/JJ area/NN seen/VBN from/INhigh/JJ PP above/IN./. The popular POS taggers could be the maximum entropy tagger (Ratnaparkhi, 1996) and the Brill tagger (Brill, 1994). For POS tagging, our system uses the OAK System which has a method similar to the Brill tagger, has 13% fewer errors (Sekine, 2002). We used POS tags in patterns for which We developed for finding answers. Chunked part of speech is making group of words with certain parts of speech into noun

phrases, preposition phrases and verb phrases. It is also defined as a shallow parse as it is done with one pass. Ramshaw and Marcus (1995) sketches a transformation-based way of tagging chunked part of speech. This is machine learning method which learns rules for whether a word belongs in a noun phrase, verb phrase, or preposition phrase, given the part of speech tags of the word and the Before tagging with chunked part of speech, the sentence looks like: After the announcement ceremony at the Smithsonian, Mr. Cling traveled to Baltimore where he announced a project to repair outdoor monuments in Baltimore. After tagging with chunked part of speech the sentence will look like:[PP After/IN ] [NP the/DT announcement/NN ceremony/NN ] [PP at/IN] [NP the/DT Smithsonian/NNP ] , / , [NP Mr./NNP Cling/NNP ] [VPtraveled/VBD ] [PP to/TO ] [NP Baltimore/NNP ] [ADVP where/WRB] [NP he/PRP ] [VP announced/VBD ] [NP a/DT project/NN ] [VPto/TO repair/VB ] [NP outdoor/JJ monuments/NNS ] [PP in/IN ] [NPBaltimore/NNP ] ,/, [PP including/VBG ] [NP the/DT Francis/NNPScott/NNP Key/NNP monument/NN ] ./. Li and Roth (2001) make use of shallow parsing for question answering, instead of a deeper syntactic parse. They identified that in certain situations, such as when lower quality text is used for generating answers, a system using a shallow parse is more effective and flexible at answering the questions. A good example of lower quality text is when the text was not edited for spelling and grammar. We developed our system for medical documents, so it should be beneficial to have a syntactic parse. We use chunked part of speech for getting question classification and for tagging the documents. his section deals with the Design and Implementation aspects of the suggested solution. The interface is the

platform where a user interacts with a map and submits his question and receives answer from system. The user interface is significant element of a QAS. Question Answering System is the core of overall system which interacts with different components user interface, Word Net, Parser, Name entity Tagger and generates output as answer to the user query. Question Answering System (QAS) performs the task of, given a user query expressed in Natural Language (NL), retrieving its correct answer in the form of compact text from large text based documents [4], [6], and [16]. Our question answering system does extract most relevant answer to the question from the large document which contains information related to medical domain. Our approach of question answering system firstly classifies the query, then extracts the candidate answers and finally ranks these candidate answers. When the user posed the question, the system extracts the answer information from the tagged document passages in terms of named entities. From these named entities some will represent the candidate answers of the question. The other part of our system finds out the semantic relation between question and candidate answers and assigns the weights to the candidate answers and ranks the candidate answers according to the weighs which gives the best accuracy with the use of Word Net [17] before generating final answer to user query. The parser is component of compiler or interpreter which act as the natural language processing tool that does syntactic analysis thus considers the grammatical construction of sentence, it can find out which words are used as subject, verb or object , phrases [6], [37]. We used Stanford Parser, a tool which generates the Parts of Speech (POS) of each word of the inputted user query

and the candidate answers selected from documents. Word Net is generally used in many QAS as it happens to be a very helpful tool when it deals with words. The Word Net characteristics are utilized to point out the relationship existing in between words of user query and data source. Word Net is the massive lexical database consisting of English words. We used OAK, a named entity tagger which tags about 150 named entities. These entities are arranged in hierarchy. We used the Penn Treebank Part of Speech tag set [27] as this Treebank is utilized to train the Stanford Parser. Here we go through system overview which consists of external tools that are being used in our system. In our question answering system, we are using a Semantic Approach for extracting the answer from the text documents. In this method sentence are selected on the basis of concept rather of specificThe knowledge base is the main source of data in our Question Answering System. This knowledge base could be in structured or unstructured form. Generally in DBMS, type of a structured data can be easily accessed. But the large amount of information storage occurs in form of the text files which are unstructured [28]. In past, Question Answering Systems were interfaced against the DBMS. The resources present as web files are being already indexed so as to retrieve the documents containing the answer is not a cakewalk. The main handy task is to find out the answer from a text document. For building knowledge base we collected information from Wikipedia [29] and other sites which are divided into various headings which are referred as context information of the disease. The updating can be easily performed manually by taking information from net, and should be placed in the document falling under appropriate headings as mentioned

above. Noise removal from documents- Noise removal transforms the raw document into the document which comprises of only content related to subjects of the document (not contain images, header, footer and other irreverent text). The Text is the sequence of characters. Tokenization is fragmenting a string of characters into its lexical elements such as words or punctuations [16]. Sentence splitting is the process in which we do splitting of the words and punctuation into their separate sentence [16]. Both the Parser and Named Entity Tagger make use of these techniques before doing tagging of documents. This module does tagging of useful information from knowledge source to bring out the cosine information. For the specific document we used name of disease as document name. The system acknowledges different types of entities which are taken from the passage or query terms based on the rule allotted for each entity. The tool ‘ Named Entity Tagger’ does the task of tagging very efficiently in our system. The question categorization module provides expected candidate answer type of the question. If tagging of the named entity is done successfully, final candidate answer can be easily generated by the answer extractor module. This sentence is selected as the candidate answer from the above paragraph as it contains the time. We have used efficient named entity tagger OAK [27] which can do tagging of about 125 named entities. All the 125 named entities are present in hierarchy. There are other name entity tagger tool available such as Lingpipe or Stanford NER but these name entity tagger can do tagging of very limited types of name entities only which are shown below: Hence we find that OAK is comparatively better tool for doing tagging of the paragraph. OAK tools works as a rule based English analyzer that has

many functionalities (POS tagger, stemmer, Named Entity (NE) tagger, dependency analyzer, parser, etc). The parser is component of compiler or interpreter which act as the natural language processing tool that does syntactic analysis thus considers the grammatical construction of sentence, it can find out which words are used as subject, verb or object, phrases [6] , [37]. Following figure 3. 2 shows parser functions. Grouping of each word is done with its parts of speech (POS). Tagging parts of speech of the words depend on the way whether the word is adjective, noun, or verb. Generally natural language processing does additional POS tags such as ‘ noun-plural’. The part of speech gives syntactical information about each word which is being used as the noun or verb. For example, the word Nice can be used as adjective (He is very nice) or as noun (Nice is a good medicine). We can do tagging manually of a document but it is really very time consuming process, there are many automated systems available which can do tagging with fairly high accuracy. For this task we have Stanford Parser, a tool which generates the Parts of Speech (POS) of each word of the inputted user query and the candidate answers selected from documents. To get semantic relation for similarity check between each words of query with each word of selected candidate answers Word Net is the most efficient tool. It is used in most question answering system because of its success when it comes to deal with words. Hence Word Net is used to discover the relation between query words and document words. It is a massive lexical database consisting of English words. These are referred as cognitive synonyms (synsets). Synset has a collection of words which are having same meaning otherwise it does grouping of different senses of word placed in different synsets. Synsets are

linked together through semantic-conceptual and lexical relations. Word Net keeps the track of each word which does similarity check. Each sense of word is confined to a synonym set. Word Net is constructed ontology which is having metonymy, synonym, and hyponyms relations [9]. So we see in most of the cases where the words of inputted query does not exactly matches in document sentence pattern , so to deal with such case we are using Word Net to find the semantic relation between query words and documents words take another example: For this we do plain pattern matching for classification of inputted question into these categories. There is another sub classification done to give more accurate expected answer to the question. There is supervised machine learning proposed by the Li and Roth [45] and Hermjakob [46] also perform classification of the questions into these categories. But comparing between plain matching methods and supervised machine learning we found that manually classification done in case of the question is the quite better approach as more accuracy are observed. We analyzed about the medical domain and then finally categorized it into eight categories and then further sub categorization is done [2], [5]. In most cases, Does, is, are, has comes after What, Which, When, Where etc. If the query put up by user begins with " When", it means that the expected answer will be words containing DAY, DATE and MONTH sorts of information. It thus incorporates event based information system. The Answer for question that begins with " What" requires taking the entire sentence which has important events occurred in document. Here we observed that question beginning with " What" can be classified by pattern matching algorithm along with finding out the focus of the question. Take for example " What are the

medicines for tuberculosis?" The focus of question is medicine. These focuses are then compared to the sentences which contain these named entities, and only that sentence is taken as the candidate answers of the inputted question. So it is very crucial to find out the focus of the question which gives hint about what sort of entity is going to be present in the answer. List of questions, classified by theirOur system does answer the question which starts with " Who". Who-definition questions passes goal directly to the answer extraction module which generates useful information about the goal. Similarly the answer to the Who-List consists of named entity more than one person. For Who-Type of Factoid Questions the Answer extraction module does extracts the facts about name entity that can be person or place. In the medical domain there are a few question starting with " Who" Such asThe Answer for question that begins with " Which" requires to find out focus of the question. In both these question focus is different i. e. in first one focus is on searching for the named entity medicine, whereas in second one focus is on searching for the names entity disease. Thus the sentence containing respective named entity is taken as the candidate answers of the inputted question. So it is very crucial to get the focus of the question which gives hint about what sort of entity is going to be present in the answer. The question beginning with " Where" is very significant in medical domain information system. Our system does answer the question which starts with " How" quite effectively. The Question that begins with " How" are divided into many sub-categories which are outlined in Table 3. We hereby discuss the method of generating the candidate answers for the inputted different types of questions in this section. In this regard the Answer

Extraction module does extraction of the candidates answer, and further does ranking of the candidate before finally presenting it to user[3],[13]. Then Answer validation is performed. Answer validation is also a Key part of thisIn the previous section we dealt with extraction of the candidate answer from the document. Answer Ranking module perform ranking for each candidate answer and show it to user based on the ranking [15]. This answer ranking module should only produce exact answer. Ranking is done as confidence estimation which has been discussed by Xu et al [13]. This ranking is based on the semantic relation who calculates similarity between the question and the candidate answers with the help of Word Net. Word Net is the vocabulary of words in which words along with their semantic meanings are arranged in hierarchical tree.