

**Project Status Report**



**Project Name:** Q&A system (Natural Language Processing)

**Department:** SOCIT

**Focus Area:** Medical Domain ( Diabetes)



**Prepared By: Espiritu,Immanuel Tan**

|  |  |
| --- | --- |
| **Document Owner(s)** | **Project/Organization Role** |
| Espiritu, Immanuel T. |  |
| Febrer, Jayvee D. |  |
| Amata, John Patrick |  |

# PROJECT STATUS REPORT PURPOSE

As of now we have already finalized our domain of knowledge where we will be focusing on Question and Answering system that is Health related, specifically Diabetes. Our target is to re-study an existing paper and find a way to scrutinize or prove if its theories can still be applied now a day’s, regarding existing systems and technologies. We collected 100 questions to serve as our material for our Q&A system. We need these questions to be formatted in a certain model our reference paper suggested.

We are now currently studying different existing papers, here are summarize versions of some papers we are currently reading:

**New Trends in Automatic Question Answering:**

Christian Gailer ([christian.gailer@student.tugraz.at](mailto:christian.gailer@student.tugraz.at))

Stefan Kohl ([stefan.kohl@student.tugraz.at](mailto:stefan.kohl@student.tugraz.at))

Stephan Oberauer ([stephan.oberauer@student.TUGraz.at](mailto:stephan.oberauer@student.TUGraz.at))

(Group Assignment Information Search and Retrieval Graz University of Technology WS 2012/2013)

The following paper focuses on Automatic Question Answering (Automatic QA), a sub- field of Information Retrieval. The first chapters focus on historical developments and the definition of the field. Thereafter, an overview of current research topics and key aspects will be given, as well as a classification of the most interesting approaches. The main part of this paper is an analysis of new trends regarding Automatic QA, primarily focusing on approaches in association with web and new media technologies. A discussion of available tools will follow up. Finally, a summary of the things learned during the research on the topic will conclude this work.

In the beginning of Automatic QA the studies where in a shape of creating an intelligent computer system, which can interact with a human being. It evolved from simple interaction systems without a knowledge database relying on a domain specific field to complex systems, which are web-scaling and able to answer elaborate questions in an interactive and context based wa

**Structure of the Work**

This paper focus on new trends in Automatic QA. The Automatic QA Process can be divided into 4 major parts: Question Analysing, Preparing the Dataset, Text Processing and Data Mapping.

The Information Retrieval part in the Automatic QA project takes care of finding documents, which contain useful information for the question answering. As it was common in the previous years that in Automatic Q&A only document retrieval was used, the information retrieval approach nowadays can also deliver text passages where valuable information can be extracted.

The **Natural Language Processing** approach performs a semantic analysis of text. It therefore uses machine learning algorithms to learn rules for text analysis. It uses sets of theories and technologies. In the early years the most common algorithms were decision trees, the latest approaches base on statistical model and probalistic decisions.(Gunawardena, Lokuhetti, Pathirana, Ragel & Deegalla, 2010)

The first kind of Automatic QA system is so called ELIZA. The system relies on natural language processing and formed a question out of a statement by using simple pattern matching. It worked on a MAC time-sharing system at MIT. The program could run different scripts to simulate a human conversation partner.

**Computational Linguistics**

|  |  |
| --- | --- |
| Maximum Entropy Model | The basic idea of the maximum entropy model is, that the probability distribution which best represents the current state of knowledge is the one with the largest entropy. It is used to build up models of many different sources with limited information. |
| Decision Tree Learning | Decision tree learning is a method used in many domains of knowledge discovery, pattern recognition and data mining. Decision trees are hierarchical trees which try to predict an output variable for a given input. Each leave represents an attribute. Each path is a conjunction of the attributes which are on the path. |
| Artificial Neural Networks | Artificial neural networks are mathematical models based on the biological neuron structure of the brain. The neurons, also called nodes, together with the weighted connections are the basic components. The advantages of artificial neural networks are the possibility of enhancing the processing speed by parallelization, adaptation of knowledge, robustness and implementation in low power applications |

**Information Extraction**

|  |  |
| --- | --- |
| Entity extraction | It identifies and classifies all phrases in a free text which refer to objects of semantic classes like names, nouns, pronouns etc. In addition all object mentions are linked together which refer to the same entity. |
| Relation extraction | The relations between entities are identified. A relation is always represented by two entities and can be described in many languages. |
| Event extraction | It is also a common application in order to derive specific knowledge from a text. Event extraction identifies events of particular types and the corresponding arguments. A type of an event would be for example ”car crash” or ”natural disaster”. |

**Architecture of an Ontology-Based DomainSpecific Natural Language Question Answering System**

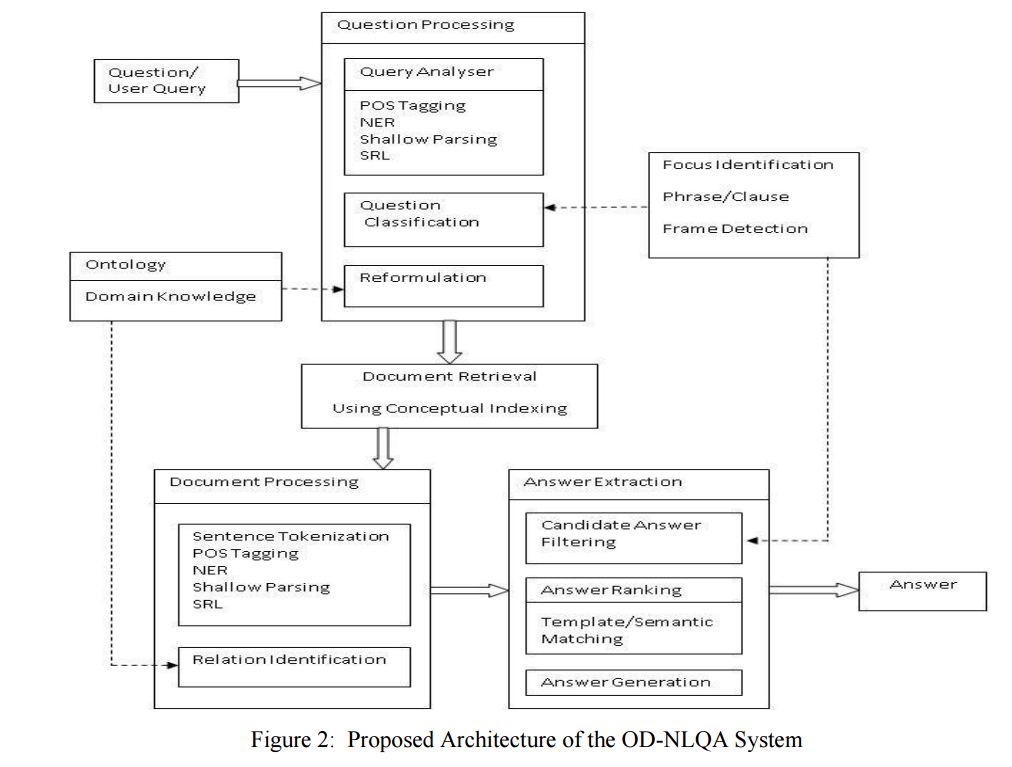
***Athira P. M., Sreeja M. and P. C. Reghuraj***

*Department of Computer Science and Engineering, Government Engineering College, Sreekrishnapuram, Palakkad Kerala, India, 678633*

“Question Answering, the process of extracting answers to natural language questions, is profoundly different from Information Retrieval (IR) or Information Extraction (IE). IR systems present the user with a set of documents that relate to their information need, but do not exactly indicate the correct answer. In IR, the relevant documents are obtained by matching the keywords from user query with a set of index terms from the set of documents. In contrast, IE systems extract the information of interest provided the domain of extraction is well defined. In IE systems, the required information is built around in presumed templates, in the form of slotfillers.”

**THE PROPOSED ARCHITECTURE**

The proposed architecture of an ontology-based domain-specific NLQA system is depicted in Figure 2. The model integrates key components such as Natural Language Processing techniques; Conceptual Indexing based Retrieval Mechanism, and Ontology Processing.

****

**Question Processing**

In the question processing module, with the help of various components, the following actions are performed.

· Analysis of the natural language question

· Question classification

· Reformulation of the user query

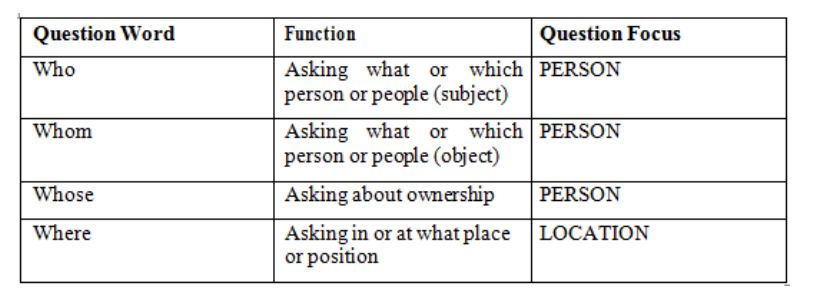
**Query Analyzer**

The natural-language question given by the user is analyzed using various natural language processing techniques.

*Syntactic Analysis* – The question is analyzed syntactically using NLP techniques. Part-ofspeech tagging and named entity recognition (NER) are performed. Tools such as Pythonnltk, OpenNLP, Stanford CoreNLP can be used for this purpose. In the proposed system, we used Stanford CoreNLP tool-kit. The CoreNLP processes the document and creates an XML file as output. Shallow parsing is performed to identify the phrasal chunks. The phrasal chunks can be identified using the Regular-expression chunker and the Conll-2000 trained chunker.

*Semantic Analysis -* Semantic role labeling is an important step in this module, which enables to find the dependencies or restriction that, can be imposed, after getting the user query . This greatly eliminates the chances of irrelevant set of answers. Semantic roles are identified using the verbnet frames.

**Question Classification**

****

**Query Reformulation**

The user queries may be reformulated by adding domain knowledge and ontological information.

**Document Retrieval**

This module selects a set of relevant documents from a domain specific repository. Conceptual indexing is used for the retrieval process since the key word based indexing ignores the semantic content of the document collection . Both the documents and queries can be mapped into concepts and these concepts are used as a conceptual indexing space for identifying and extracting documents.

**Document Processing**

The retrieved documents are processed for extracting candidate answer set. This module is responsible for selecting the response based on the relevant fragments of the documents.

|  |  |
| --- | --- |
| **Syntactic Analysis** | The documents analyzed syntactically using the NLP techniques such as part-of-speech tagging and named-entity recognition. |
| **Semantic Analysis** | Shallow parsing can be performed for finding the semantic phrases or clauses. The semantic roles are identified and mapped to semantic frames. The sentences whose semantic frames map exactly to the semantic frames of the question are also extracted. |
| **Relation Identification** | The base ontology is populated with the domain knowledge incrementally as we go through different set of documents. By this method a valid knowledge on any specialized discipline can be incorporated to the system. The relations among different concepts are identified using the domain knowledge and the ontological information obtained. |

**Answer Extraction**

The filtering of candidate answer set and answer generation is performed. The user is supplied with a set of short and specific answers ranked according to their relevance. The different stages are, *Filtering*, *Answer Ranking* and *Answer Generation.*