International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org/journals.php
Vol. 7 Issue 9, September-2017, pp. 27~34
ISSN (O): 2249-3905, ISSN(P): 2349-6525 | Impact Factor: 7.196

Information Retrieval from a Structured Knowledgebase

Dr.Poonam Yadav
Assistant Professor, D.A.V College of Engineering & Technology,
Kanina, Haryana 123027, India

Abstract

There are a vast amount of data available on the internet. This makes the process of searching difficult for the user. The user struggle a lot to go through the entire datas and find the appropriate information. Hence it is necessery to develop an automated system that makes searching easier. This paper proposes the the method of Question Answering(QA) which makes the searching process easier. The motive of QA system is to provide the relevant answers for the questions raised by the user. The process made to select the precise answers in QA is becoming more challenged when compared with search engines. The motive of this paper is to create an easier way to find answers based on relavant questions by using Information Retrieval (IR) and Natural Language processing (NLP). Also it proposes the use of annotated and structured knowledge rather than using unstructured language. DBpedia is used as knowledge base and TREC 2004 dataset is used for final representation.

Keywords: Information Retrieval, Natural Language Processing,Question Answering System,Structured Semantic Data.

INTRODUCTION

Question Answering (QA) is a process of answering to a question which is asked in a natural language by the user. The QA system's main purpose is to generate information containing in a knowledge based query and answer it according to the question asked by the user. It can be either unstructured (documents in English) or structured (knowledge base or database). QA is of two domains: 1. open domain and closed domain QA systems.

Motivation

It is difficult for the user to navigate all information's as the data amount in internet is uncountable. Hence nowadays several researches are going on to enhance the retrieval of data quickly. Frequently Asked Questions (FAQ's) are asked in a large number on the internet. Two challenges are faced while using this method. 1. Explicitly, the questioned cannot be asked. 2. Each question should be passed through to find the matched query. Thus it will be an unintuitive and time consuming method (Rao Yerrapragada et al., 2014) Even though web search engines help to answer the questions it is necessary for the users to find the correct information. Querying the information availed in databases is another form of searching. SQL language should be known to those who prefer this method. Hence human language is the best way to search for a normal user. Also using this method, the user can find out the exact answer for his/her question.

LUNAR (Woods et al., 1972) that provides answers to question based on rocks and their geological analysis to Apple's Siri (Roush, 2010), a personal assistant which is an intelligent one. All these systems are to provide the users with the answers present in their data bases they are searching for the big problem is to verify and decide the right answer. The reason behind it is the huge availability of data on the internet. The knowledge format used is important in the QA system**.** The returning of information into specific pieces of tasks can be obtained from an IR perspective. For providing long type of answers this is considered as a better approach but for providing the exact

answers it cannot be preferred. NLP methods check the semantic and syntactic structures and try to understand it. Named Entity Recognition is a method used to recognize the entities in NLP method (K. S. S. Rao Yarrapragada and B. Bala Krishna, 2016). The information retrieval is facilitated by an advanced compression approaches (B.S. Sunil Kumar et al., 2016) and modern data transmission technologies (K Bhatnagar and SC Gupta 2017) and (Kavita Bhatnagar and S. C. Gupta 2016). Retrieval methods and natural language (P. Vijaya, and Satish Chander 2016) makes a combination and make a difficult approach to answer the questions (S Chander et al., 2016). Structured database has information that is additional and helps to find the information in an easy way.

Problem Statement

The main motive of this paper is to create a novel that will be robust to QA on the basis of natural language processing and retrieval of information. It is analyzed using a structured database. The specific questions are (i) does it have certain knowledge aid based on structure? (ii) Whether the annotation format matters when it is knowledge based.

RELATED WORK AND BACKGROUND

Question Answering Systems

QA system is widely used for its results and applications. It aims to provide the user with accurate answer for their questions. Information retrieval, Natural language processing and Information extraction are combined to form a QA system. Starting from BASEBALL (Green et al., 1961) to Siri (Roush, 2010) there have been a large number of breakthroughs. The accurate results and demands have motivated research and works in QA field. It can be divided into two types. 1. Open domain QA and 2. Closed domain QA. The communicating medium plays a vital role in recognizing the QA systems (P Yadav and RP Singh 2012). It can be voice controlled i.e. when the question is spoken by the user; the system gives the answer back in natural language or else on a text basis, where the question and answers are in textual format. Examples of voice based QA systems are Google and Siri (Roush, 2010). Examples of text based QA systems Wolfram Alpha (Wolfram, 2007), START Katz, 1997). From 1995 (Voorhees, 2004), Text Retrieval Conference (TREC) provides a track for question and answering, i.e. to answer questions for short factoid questions. Factual questions can be answered by a QA system. Now researches are made to provide answers for complex questions. If a question starts with 'why' it will be more complex.

Related Work in QA systems

QA system has three steps: Information Retrieval, Question classification and Answer extraction. A mapping technique is used by Ruger and Cooper (2000) that classifies the interrogative type of question to indicate the class and focus of the question. The questions that begins with why, whom, who, when, where can quickly be found out using direct mapping method. The questions that begin with which, what etc is given to Word Net (Miller, 1995). To determine the type of question, an additional context is given. Roth and Li used a machine learning method and briefed their aim as, "to classify the questions depending upon the answers that are on semantic types". They proposed a hierarchy of semantic answer types (Li and Roth 2006). LUNAR (Woods et al., 1972) and BASEBALL (Green et al., 1961) is the oldest model that uses natural language to query structural database. The questions are converted into a formal one to find the answers. A program called ELIZA (Weizenbaum, 1966) is said to make the conversation between computer and natural language possible for the first time. The responses made by analyzing the input statements depending upon decomposition rules are triggered on the basis of some keywords. Borchardt,

Felshin and Katz used a method to add certain structures known as Natural language annotations that matches the questions and answers. The knowledgebase and resources of information are added manually. It is a new technique by which information resources can be indexed. McGowan interprets the problem of retrieval which converts the questions in to formal search (Mcgowan,2016). Barskar et al. interpreted on pattern learning based on extraction that focus to provide answers which are "natural and complete" (Barskar et al., 2012). Finally, QA system has improved its performance from earlier decades. From this approach, we can conclude that the researchers were heterogeneous in QA fields. In this paper, we look in to certain techniques of Natural Language Processing and Information Retrieval.

QUESTION AND ANSWERING IN A PROPOSED SYSTEM

The QA model has three steps that are explained in this section,
1. Questionprocessing
2. Document processing information retrieval
3. Answer extraction. .
The architecture in high-level is shown in Fig. 1.

Knowledgebase

QA method prefers for a "resource" to provide answers. Document corpus, World Wide Web or data bases etc are examples for resources used by QA. It is divided into two types 1.Restricted domain question answering (RDQA) and Open domain question answering (ODQA). RDQA method uses knowledge resources from databases where knowledge is encoded. From earlier databases, it can get the answers. Thus it mainly aims on information's that are domain specific to query (Mollá and Vicedo, 2007). Here the knowledge resource is incorporated in QA model after being built up by the person who are experts in domain region. QA track in TREC (Voorhees, 2004) leads to development in open domain QA. WWW (World Wide Web) is used as resource knowledge in ODQA which uses a text database that can be larger. Knowledgebase is divided into 3 types. 1. Structured 2. Semi structured 3. Unstructured.

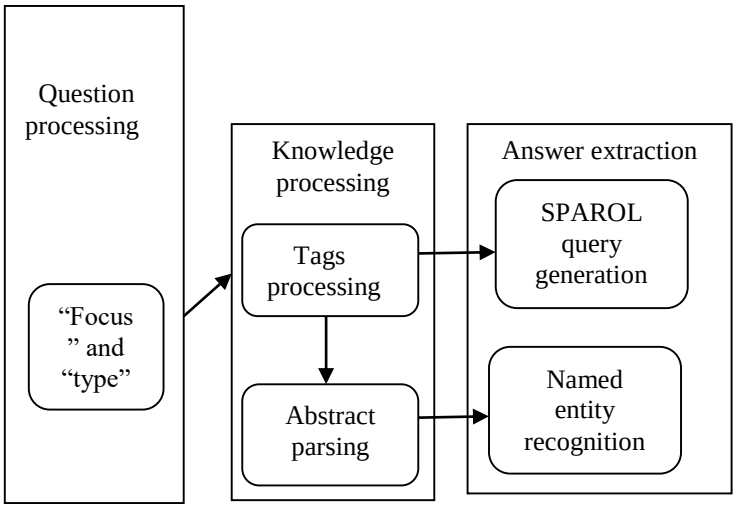

Fig- 1: The overview of high level system

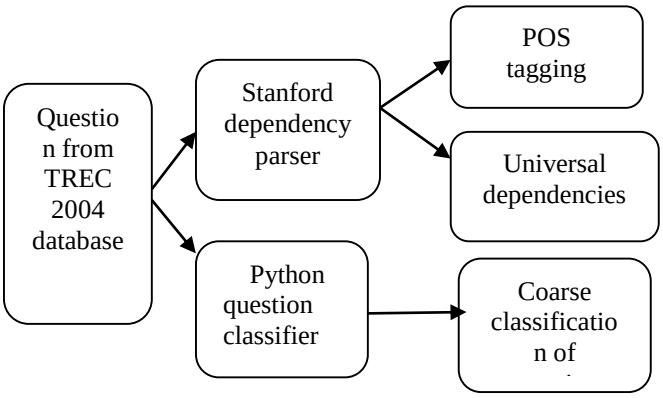

Fig- 2: Processing of question

Processing of Questions

A sentence that starts with an interrogative word is defined as a question. To provide answers, it is necessary for the system to understand the question. Hence the question should be progressed, parsed and tagged first. Fig.1 implements the QA flow. The question is provided into a method developed by Ruth and Li called Python factoid question classifier that provides the right answer for the question asked. Then using part-of-speech and Stanford Dependency parser, tagging is obtained. Verb and noun plays a vital role in identifying the question. Not only POS tags, but also the dependencies of Stanford universal are obtained. Thus parser and classifier create a sense on the question.

Question Classification: constraint determination is a important process in a QA system. Filtering unnecessary answers is also important to ensure accuracy. Question classification (QC) is used to classify and identify the type of question easily. Let us assume a question "where was Obama born?" for this type of question the QC identifies all the locations and removes all other type of questions. It has two main processes. 1. To place constraints. 2. Giving further information for selection process.

Classification based on Python Factoid Question: machine learning system is used by Python factoid to classify the questions. It classifies the question and answers based on varied semantic classes. The question that possesses an action is not classified using this method. The questions that starts with "why", "where", "when", "who", "which", "what" are only addressed. NUMERIC VALUE, LOCATION, HUMAN, ENTITY, DESCRIPTION and ABBREVIATION are the six coarse classes which consist of a class of 50 non-overlapping sets (Li and Roth 2006).

Stanford Dependency Parser: It is a java implemented method and to parse the question, Core NLP toolkit is used in Stanford method. It provides functions for syntactic parsing and speech tagging. The verb and noun indicates the type of question asked. Let us assume the earlier question, "where was Obama born?" It provides a clue that the question is indicating the birth of Obama. Thus syntactic parsing is the second most necessary process in QA technique. Universal dependencies are relations of grammar that returns a parser in a line. The frequently used processing modules are analyzed by standard dependency parser and QC.

Knowledgebase Processing

DBpedia is used to retrieve data that is matching to the question in this module. Abstract Parsing and Tags Processing are the main factors of this module. Key is denoted by tag or label and in DBPedia, the pair of key value represents the data. It is developed manually that can be given as a

simple annotation manually and is assigned to each value. Thus tag processing is the primary step to be followed to find the correct answer for the particular question. Hence from DBpedia, all the suitable tags are collected, classified and ranked accordingly. Based on feature matching, the tags are ranked. For that a ranking algorithm is used. On the basis of the amount of features matched, scores are assigned by the ranking algorithm. By using SPAROL query method, the answer is derived from DBpedia on the basis of tags that are ranked higher. The module used for abstract parsing is employed when there are no matches for the particular question. It also has a short and long abstract for each page. The abstract that is short is retrieved and then parsed. To select an appropriate answer, pattern matching techniques and Stanford Universal Dependencies are used. For extraction purpose, simple heuristic rules and Named Entity Recognition are employed.

Answer Extraction

Named Entity Recognition and SPARQL query generation are the techniques used to extract the answer. After the processing of tags, the tag which is scored higher (no zero score) is given to the SPARQL query generation module. The Abstract parsing module holds the control if zero becomes the largest score whereas, the Named Entity Recognition sub-module holds the control if the highest rank is not zero. Extracting answers is a part of QA process that provides the correct answer from the collected information (Wang, 2006). In the earlier method, the suitable answer is obtained from an appropriate tag and then processed. The function of the answer extraction module is to obtain the right answer according to the given information. IR paradigm is used by the SPARQL query generation to generate a query and get the answer in a language that is formal. At the same time, NLP paradigm is employed to operate and parse a sentence in a language that is natural.

IMPLEMENTING WEB APPLICATION

This method is employed on Apache Tomcat local Server and it has a Java backend. It provides executions and arguments by communicating with QC. For querying purpose in DBpedia, a framework called Apache Jena is enhanced. JavaScript and HTML is used to build a user interface here. In Python, Factoid QC (Li and Roth 2006) is employed.

Results

The proposed QA method was based on TREC 2004 dataset (Voorhees, 2004). It is aimed to motivate research in the field of QA systems. The datasets in TREC systems has questions based on short answers and fact based questions. The dataset in TREC 2004 has a number of questions that are formulated on the basis of a target. It can be a thing, organization or a person. The questions in the dataset ask for more details relying the target. The targets and the questions raised earlier give a context to the present question and so the order of the questions also carries much importance. The dataset can identify whether the questioner is an adult trying to encounter relevant information or basically is an English speaker. There are 65 targets in the TREC 2004 dataset where the number of organizations is 25, the number of things is 17 and the number of people is 23. 286 questions are there in Factoid TREC 2004 dataset.

In fig. 3, the TREC 2004 dataset is explained where a person is targeted in 22nd series, organization is targeted in 21st series and thing is targeted in 3rd series. In XML document, the target that are associated and the questions are encoded.

3. Hale bopp star 3.1 FACTOID: When was the star discovered? 3.2 FACTOID: How often does it approach the earth? 3.3 LIST : In which countries the star is visible on its last turn? 3.4 OTHER
21.Museum 21.1 FACTOID: How many museums are there in France? 21.2 FACTOID: Where is the deer museum located? 21.3 LIST : List the number of museums in India? 21.4 OTHER
22. William Shakespeare 22.1 FACTOID: Where was William born? 22.2 FACTOID: When was he born? 23.3 FACTOID: What is his background? 24.4 LIST : What are the books written by him? 25.5 OTHER

Fig- 3: Shows the sample question series available in TREC 2004 dataset

Table 1 and 2 represents the results based on its accuracy level after being tested against the TREC 2004 questions. The overall accuracy is given by Table.1. It classifies whether the given answer is right or wrong and denotes the breaking down of the question. The incorrect answers are denoted by Table 1 and it breakdowns the questions. Table 2 uses the two tracks and classifies the right and wrong answers, i.e. abstract parsing and tags processing.

Table 1- Accuracy in terms of abstract processing and tags processing

Number of questions answered incorrectly or not answered.	Number of questions answered incorrectly	Number of questions whose answers are not present in kb
161	68	93

Table 2- The questions that are not handled by the system are broken down

Module	Number of questions answered incorrectly	Number of questions answered correctly
Abstract processing	40	26
Tags processing	22	99

Analysis

The research questions are again visited to analyze the results. The primary question raised was "whether the availability of structured knowledge base helps in improving a QA system?" The answer is 'yes' and it can be derived by using this prototype. The data can be retrieved quickly by using a semi structured or structured database. The system offers hints or clues so that the answers can be obtained easily. It also reduces the searching time and thus processing time is reduced to a great extent. It shortens the searching process just by indicating the certain paragraph or location. The method for Structuring of knowledge or database is the next procedure to be used by the system. A large number of ways are there to structure the knowledgebase but among them 'annotations' is the well known and very familiar technique. Most often a question is raised that whether the annotation format is considered during the process of structured knowledge base. Now the answer can be given as 'yes'. The annotations and tags are in DBPedia are formulated by various individuals manually. Hence the annotations and tags cannot be unique

since each individual follows his own style of designing the annotations. The tags and annotations can differ from each other and they maybe ambiguous often. Thus the tags and annotations that are standard throughout the whole knowledgebase and prefer a certain structure or format are considered to be helpful rather than using manual annotations.

CONCLUSION & FUTURE SCOPE

The combination of works from NLP and IR is represented in this paper. The knowledge resource used determines the success of the QA method. Thus by using an annotated knowledge base with a structure greatly aids in improving the system and also it enhances the system performance. Many enhancements can be made to each techniques and modules. The proposed QA system relies upon the question and subject from UI (user interface). The XML document which is encoded in TREC dataset 2004 generates the UI. A textbox can be added by which users can ask their own questions that can make the system more enhanced. To extract the 'target' or 'focus' from a question, another method called Stanford Dependency parser can be used instead of using DBPedia which depends on the usage of targets. A small challenge is that it depends on Python QC (Li and Roth 2006). Almost 20% of the answers are incorrect because of the wrong classification made by the Factoid QC. The QC which is further trained better can be used for improving the system. While parsing, addition of semantic role labelling module, also known as shallow semantic parsing will also enhance the performance. It allocates certain functions after indicating the semantic arguments linked with a predicate or verb. Shallow semantic parsing is designed to provide varied sentences with standard and formal representation with same meaning. It also minimizes the ambiguity level. Other main reason for the wrong answers is the insufficiency of knowledge resource in DBpedia. This can be rectified by using certain structured knowledge resources such as NELL or YAGO. These can provide the answers for those questions which are not in DBpedia. Word2vec is a semantic parsing tool and is helpful in abstracting parsing module. Using such semantic tools provides good results in understanding the languages than using syntactic parsing method.

REFERENCES

Allam,A.M.N, and Haggag, M. H. (2002): "The question answering Systems: A survey", In Intl. Journal of Research and Reviews in Information Sciences (IJRRIS), 2(3): 211–220.
Barskar, R., Ahmed,G.F., Barskar, N. (2012): "An approach for extracting exact answers to QA system for english sentences", In Procedia Engineering, 30:1187–1194.
Bizer, C., Lehmann,J., Kobilarov,G., Auer,S., Becker,C., Cyganiak, R., Hellmannb, S., (2009): "DBpedia - A crystallization point for the Web of Data", In Journal of Web Semantics, 7(3):154–165.
Cooper, R., and Ruger, S, (2000): "A simple question answering system", In Voorhees and Harman.
Green B.F., Wolf. A.K., Chomsky, C., Laughery,K. (1961): "Baseball: An automatic Question Answerer",In Proceedings of Western Computing Conference, 19: 219–224.
Katz. B, (1997): "Annotating the World Wide Web Using Natural Language", In Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet.
Kolomiyets, O., and Moens, F. (2011): "A survey on question answering technology from an information retrieval perspective", In Information Sciences, 181(24): 5412–5434.
Li, X., and Roth, D. (2006): "Learning question classifiers: the role of semantic information", In Natural Language Engineering, 12(3): 229.
Manning, C. D., Surdeanu,M., Bauer,J., Finkel,J., Bethard,S.J., and McClosky,D.(2014): "The Stanford CoreNLP Natural Language Processing Toolkit", In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 55-60.

Mcgowan,K.(n.d.).“Emma,”A Natural Language Question Answering System” from www.umich. edu.
Miller, G, (1995): ” WordNet: a lexical database for English”,Communications of the ACM, 38(11): 39–41.
Mollá, D., and Vicedo, J. L. (2007): “Question Answering in Restricted Domains:An Overview”,In Computational Linguistics, 33(1): 41–61
Roush.W. (2010): “Siri, Apple’s New Old Personal Assistant App, Points Toward A Voice - Activated Future”.
Voorhees, E. M. (2004): “Overview of the TREC 2004 question answering track”, In Proceedings of TREC, http://doi.org/10.1016/S0306-4573(99)00043-6.
Wang, M. (2006): “A Survey of Answer Extraction Techniques in Factoid Question Answering”.
Weizenbaum, J. (1966): “ELIZA — A Computer Program For the Study of Natural Language Communication between Man And Machine”, In Communications of the ACM, 9(1): 36–45.
Wolfram, S. (2007): Today, Mathematica Is Reinvented. http://blog.stephenwolfram.com/2007/05/today-mathematica-isreinvented/ [Retrieved May 2016].
Woods, W., Kaplan, R., and Nash-Webber, B. (1972): “The Lunar Sciences Natural Language Information System: FinalReport”, BBN Report 2378.
K. S. S. Rao Yarrapragada and B. Bala Krishna (2016): "Impact of tamanu oil-diesel blend on combustion, performance and emissions of desel engine and its prediction methodology", Journal of the Brazilian Society of Mechanical Sciences and Engineering, pp. 1-15.
Rao Yerrapragada. K. S.S, S.N.Ch. Dattu .V, Dr. B. Balakrishna (2014):"Survey of Uniformity of Pressure Profile in Wind Tunnel by Using Hot Wire Annometer Systems",International Journal of Engineering Research and Applications, vol.4(3), pp. 290-299.
Kavita Bhatnagar and S. C. Gupta (2016): "Investigating and Modeling the Effect of Laser Intensity and Nonlinear Regime of the Fiber on the Optical Link", Journal of Optical Communications.
K Bhatnagar and SC Gupta (2017): "Extending the Neural Model to Study the Impact of Effective Area of Optical Fiber on Laser Intensity", International Journal of Intelligent Engineering and Systems, vol.10.
B.S. Sunil Kumar, A.S. Manjunath, S. Christopher (2016): "Improved entropy encoding for high efficient video coding standard", Alexandria Engineering Journal, In press, corrected proof.
BSS Kumar, AS Manjunath, S Christopher (2018):" Improvisation in HEVC Performance by Weighted Entropy Encoding Technique" Data Engineering and Intelligent Computing.
S Chander, P Vijaya, P Dhyani (2016):"Fractional lion algorithm–an optimization algorithm for data clustering”, Journal of computer science.
P. Vijaya, and Satish Chander (2016):"Fuzzy Integrated Extended Nearest Neighbour Classification Algorithm for Web Page Retrieval", Proceeding of the International Conclave on Innovations in Engineering and Management.
P Yadav and RP Singh (2012): "An Ontology-Based Intelligent Information Retrieval Method For Document Retrieval", International Journal of Engineering Science.