ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

Answer Extraction Techniquefor Question Answering Systems

Waheeb Ahmed[1], Dr. Babu Anto P[2]

Research Scholar, Department of Information Technology, Kannur University, Kerala, India[1]

Associate Professor, Department of Information Technology, Kannur University, Kerala, India[2]

ABSTRACT: We develop and evaluate an answer extraction module for Arabic Question Answering(QA) systems. Answer extraction (AE) module aims at returning only those passages of a document that exactly answer a given user question. In traditional Information Retrieval(IR) systems return a set of documents (retrieved as a response to a user query) rather than direct answers to queries or questions. Our AE is more promising than existing information retrieval and information extraction in that the returned results are single words, phrases, sentences or passages but not entire documents. Therefore, It can be used effectively in developing better Arabic QA systems and the returned answers are not extracted from a knowledgebase or structured data but from unstructured text and this requires complex Natural Language Processing(NLP) and improved IR techniques. The current version of our AE will be able to derive the logical form of the answer from the text . User questions are also converted into logical forms for easy processing by the IR module. The final answer extracted by our AE module is precise and does not contain irrelevant text. We used the Mean Reciprocal Rank(MRR) for evaluating the performance of our proposed method for AE. The obtained results shows the effectiveness of our method.

KEYWORDS: Answer Extraction, Information Extraction, Information Retrieval, Natural Language Processing, Question Answering Systems.

I. INTRODUCTION

Currently, the mechanisms available up to now in Information Retrieval (IR) just as the case in search engines such as Google , Yahoo or MSN - allow a user only to supply a set of keywords to the search engine which retrieve the relevant documents that (partially) satisfy a given query [1]. It is the user task to look for the answer in the retrieved documents himself. In recent years, the coupling of the web growth and the increasing demand for better information access has brought motivation and interest in QA systems [2]. One of the recent and active research fields where natural language processing technology is employed to extract useful information from text in text based question answering. QA Systems read texts, process their content, and answer freely phrased questions in natural language. This facility will be very useful in a wide range of applications, specifically if questions and texts fragments could be written in any language. These systems would be the best solution to the problem of vast amount of information in the era of the World Wide Web. However, as per the current situation today (and will continue for a long period of time to come), such systems can be employed only in slightly small domains, for very small portions of text, and with high development costs. When it comes to the matter of processing larger amounts of texts there have been only two techniques until now: Information Retrieval and Information Extraction. However, both techniques have real drawbacks. Traditional information retrieval (IR) techniques allow users to submit different queries over very large collections of document (several gigabytes in size) in many domains but they usually retrieve full documents (this is also true for traditional systems such as SMART [3] as well as for probabilistic systems such as SPIDER [4]). Nevertheless, this is useless if documents are dozens or hundreds, of pages long. In some cases the techniques of IR are also used to return individual passages from documents (one or more sentences, paragraphs) [5]. In these cases the number of search terms appeared in a given sentence, together with their frequency (in terms of their count and also closeness in a sentence), is used to find relevant sentences. Unfortunately, all existing IR techniques (whether applied to full documents or to separate passages) have some limitations that make them not applicable for certain important applications. First, they consider only the content words of a document (all the function words are discarded). Second, in many cases only the stem of such words is considered and used (and this stem is usually derived with the help of

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

some kind of stemmer algorithm but not a proper morphological analysis , inevitably resulting in several ambiguities). Finally, the extracted terms from the user's query are treated as isolated words where no importance for order. This applies for Boolean systems and for vector space based systems. As a result, neither model can differentiate the phrase of "information access" from that of "access information" in the first phrase "access" is a noun and in the second it is a "verb"(the ordering of information is lost), or the concept of "transfer from Russia to the US" from that of "transfer from the US to Russia" (lost function word information). In fact, several systems can use phrasal search terms (such as "information access"), to be appeared as a full phrase in the documents, but then a set of relevant documents (such as the ones containing "access information") will no longer be returned. All of these also applies for some passage retrieval systems described in the related work (such as the QA system presented in [5]). Information extraction (IE) techniques do not have the same drawbacks. They are similar to IR systems in that they also are appropriate for handling very large collections of text(of actually unlimited size, such as flow of messages) covering highly wide range of topics. However, they are different from IR systems in that they not only recognize certain messages in such a stream (those messages that comes into a number of specific topics) but also extract from those messages highly specific content data. On the other hand, IE systems do not allow for arbitrary questions (as IR systems do). However, there is an increasing need nowadays for systems that are able of finding information in texts of moderate size but which should show very high percentages of precision and recall and which should furthermore allow questions phrased arbitrarily. Moreover they should be able to deal with documents written in syntactically unstructured text. One of the current requirements is a system that extracts the exact phrase(s) in a document (collection) from whose meaning we can derive the answer to a particular question. This is the main idea behind Answer Extraction (AE). Our proposed method is developed for the purpose of extracting answers to natural language questions and the answer could be a single word, phrase, sentence or passage based on the class of the question.

II. METHODOLOGY

Since the Answer Extraction component should be able to deal with unstructured text, it needs an efficient tokeniser, and some method of dealing with ambiguities, a module that can even have the ability to analyze the text structures(syntactical analysis) and extract the meaning too(semantic analysis), and a search engine able to use the resulting knowledge base. In the following we will describe merely three components of the system: question processing, document retrieval, passage retrieval and AE module. The following sections explains the functions of these modules. Some work has been done on passage retrieval which introduce techniques for retrieving relevant passages but not extracting the exact answer from these passages[6][7][8][9].

A. QUESTION PROCESSING MODULE

In question processing module the question will be classified to determine the expected answer type. We use Support Vector Machines(SVM) which is based on a training model from our previous work[10]. The question will be given to the classifier, the output of the classifier will be a label based on the question class. The following table shows the different classes based on Li & Roth question classification model[11].

Table 1. Question classes

Coarse grain classification(Level 1)	**Fine grain classification(Level 2)**
HUMAN	Group Individual Title Description
LOCATION	Country State City Mountain other

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

NUMERIC	Count Date Money Distance Speed Percent Other
DESCRIPTION	Definition Manner Reason
ENTITY	Color Animal Technique Planet other

For example,

Question 1: "ما هو البلد الذي عاصمته تيرانا؟"
("What country 's capital is Tirana? ")
Question 1 is classified as "LOC:country". That is , the question is looking for answer of type "location" particularly "country".

Question 2:"ما هو التيتانيوم؟"
("What is titanium")
The answer type for question 2 is of type "DESC:def", i.e., the general answer type is "description" and specifically a "definition".

Question 3:"من هو مؤسس السيانتولوجيا؟"
("Who is the founder of Scientology? ")
Question 3 is seeking answer of type "HUM:ind", i.e., the main type of the answer is "human" and specifically "individual".

Question 4:" في أي سنة استبعدت نيوزيلندا من تحالف أنزوس؟ "
("In which year was New Zealand excluded from the ANZUS alliance? ")
The answer type of question 4 is ""NUM:date" ,i.e., the question is looking for a numeric answer of type "date".

Question 5:" ما الذي يدفع الجسم إلى الارتجاف في درجات الحرارة الباردة؟ "
("What causes the body to shiver in cold temperatures? ")
The answer type of question 5 is "DESC:reason", i.e., descriptive answer and specifically "reason". Finally, the purpose of identifying the class of the question/answer type is to direct the answer extraction module to apply the proper method for extracting the answer.

B. DOCUMENT AND PASSAGE RETRIEVAL MODULES

Second, the question will be tokenized and stop words will be removed. Stop words include conjunctions, and prepositions. Tokenization means splitting the question into individual terms. These terms are sent to the document retrieval module. The document retrieval module will retrieve the top 5 relevant documents containing the question terms. We used Vector Space model(SVM) for document retrieval[5]. Next, the top 5 returned documents are divided into passages.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

C. ANSWER EXTRACTION MODULE

The returned passages will be analyzed based on the question class.

- If the question class/label is HUMAN LOCATION, then Named Entity Recognizer(NER) will be used to extract the answer.
- If the question class is NUMERIC, the Regular Expression(RE) technique is used where a set of regular expressions are used to extract the specific numeric value which is required by the question.
- If the question class is DESCRIPTION. Then similarity measures are used to retrieve the passage which may contain the answer.
- If $x=(x_1,x_2,\ldots,x_n)$ and $y=(y_1,y_2,\ldots,y_n)$ are two vectors with all real $x_i,y_i>=0$, then their Jaccard similarity coefficient is defined as

  $J(x, y) = \frac{\sum_i \min(x_i,y_i)}{\sum_i \max(x_i,y_i)}$, (1)

  and Jaccard distance $dj(x,y)=1- J(x,y)$ (2)

  where x refers to question terms represented as a vector, and y refers to the sentence terms represented as a vector.The answer with the highest similarity value with the question is retrieved.

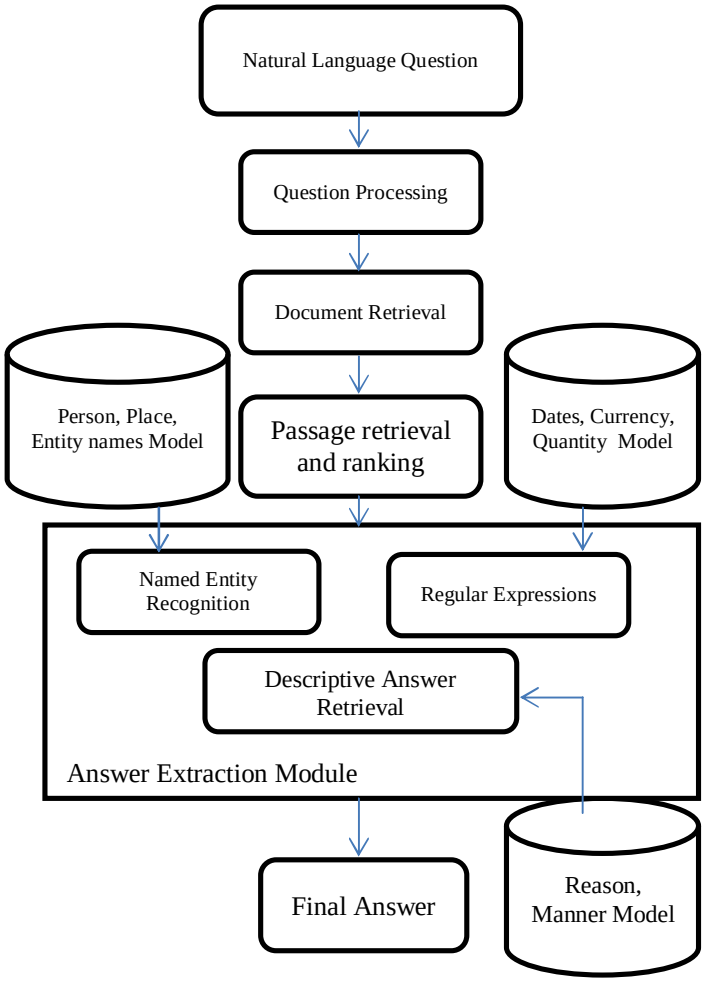

Figure 1. QA system architecture with answer extraction module

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

III. EXPERIMENT RESULTS

The experiment mainly asses the efficiency of the answer extraction method. There is no unified test data set in domain Question Answering System (QA), so this experiment uses Arabic Wikipedia as the specific domain, which selects 500 extracted documents to verify. Five question types are used to test the method which includes human, location, numeric, description and entity , each type chooses 100 questions. In the test, questions of those different types adopt different extraction strategies and methods. A total of 200 questions with 30 questions for each question type.

Table 2. Experiment results of answer extraction

Question Types	Number	MRR
		Technique (NER/RE/SIMILARITY)
HUMAN	40	.73
LOCATION	40	.68
DESCRIPTION	40	.51
NUMERIC	40	.44
ENTITY	40	.46
AVERAGE	40	.56

For the different types of questions, the assessment methods of answer extraction based on Text Retrieval Conference(TREC), using MRR (Mean Reciprocal Rank) standards shown in the following formula:

$$MRR=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{r_i} \quad (3)$$

Where, n is the number of the questions to be testedand $r_i$is the position of the first correct answer to the question number.i, if there is no correct answer available in candidate sentences, the value is Considered to be 0.

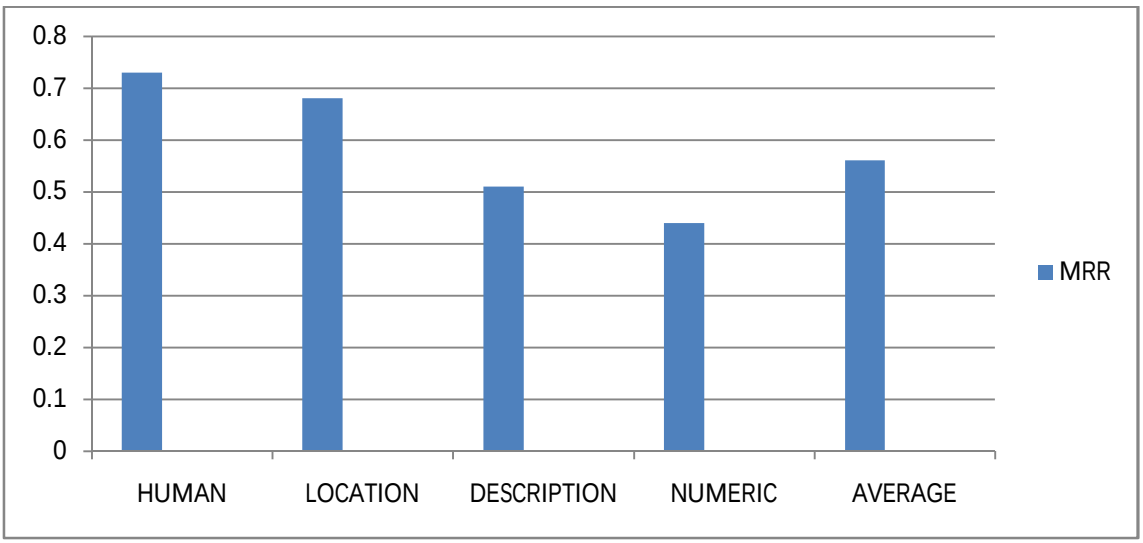

Figure 2. Distribution of the MRR for the different types of questions

Fig. 2 shows the scored value of the MRR for several types of questions tested on our method for answer extraction. With "HUMAN" question type the method performed very well and that is because the Named Entity Recognizer can identify named entities easily. However, for the NUMERIC question type, the Regular Expressions sub-module finds difficulty in identifying all formats of dates and that is because the dates sometimes do not follow the regular format

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

and comes in the form of text instead of numeric format.The overall average value of MRR for the system is considered to be very good comparing with results obtained in the literature of QA[12][13].

IV. CONCLUSION

The answer extraction is one of the key components of the Question Answering System (QA). This paper presents and evaluates a method which adopts several techniques for extracting different types of answers including human, location, numeric, description and entity by applying named entity recognizer, regular expressions or semantic similarity between sentences and question to extract answer. The experimental results shows the effectiveness of this method in extracting the answer.

REFERENCES

1. Baeza R., and Ribeiro B., "Modern Information Retrieval.", ACM Press, New York, Addison Wesley, 1999.
2. Burger J. et al. , "Issues, tasks, and program structures to roadmap research in question & answering (q&a)", In NIST, 2002.
3. Salton G.,"Automatic Text Processing: the transformation,analysis, and retrieval of information by computer", AddisonWesley, New York, ISBN:0-201-12227-8 , 1989.
4. Schauble P., "SPIDER: a multiuser information retrieval system for semistructured and dynamic data",In Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 318–327, 1993.
5. Salton G., Allan J., and Buckley C., "Approaches to passage retrieval in full text information systems",In Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 49–58, 1993.
6. Yassine B., Paolo R., and José M.,"Adapting the JIRS Passage Retrieval System to the Arabic Language", Computational Linguistics and Intelligent Text Processing, Springer,Vol. 4394, pp. 530-541, 2007.
7. G´omez J. M., Sanchis E.,and Rosso P.," A Passage Retrieval System for Multilingual Question Answering",In the 8th International Conference of Text, Speech and Dialogue 2005 (TSD'05), Lecture Notes in Artificial Intelligence (LNCS/LNAI 3658), pp. 443-450, 2005.
8. Taoufiq D., Josiane M., and Quoc D., "Passage Retrieval Using Graph Vertices Comparison", Third International IEEE Conference on Signal-Image Technologies and Internet-Based System,pp. 71 – 76, 2008.
9. Ku L.W., Liang Y.T., and Chen, H.H. ," Question Analysis and Answer Passage Retrieval for Opinion Question Answering Systems.", International Journal of Computational Linguistics and Chinese Language Processing, Vol. 13, No. 3, pp. 307–326, 2008.
10. Waheeb A. and Babu A., "Classification of Arabic Questions Using Multinomial Naïve Bayes And Support Vector Machines", In the International Journal of Latest Trends In Engineering And Technology, pp. 82-86, SACAIM, 2016.
11. Li X. and Roth D., "Learning Question Classifiers: The Role of Semantic Information", In the Journal of Natural Language Engineering, Vol. 12, pp. 229 - 249 , 2006.
12. Amine B.,Lamia H.,and Hatem G., "Question focus extraction and answer passage retrieval", IEEE/ACS 11th International Conference onComputer Systems and Applications (AICCSA), pp. 658 – 665, 2014.
13. Lahsen A. ,"On the improvement of passage retrieval in Arabic question/answering (Q/A) systems", In Proceedings of the 16th international conference on Natural language processing and information systems, NLDB'11, Springer, Vol. 6716, pp. 336-341, 2011.