ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 12, December 2016

# Answer Extraction and Passage Retrieval for Question Answering Systems

**Waheeb Ahmed, Dr. Babu Anto P**

***Abstract*—Question Answering systems (QASs) do the task of retrieving text portions from a collection of documents that contain the answer to the user's questions. These QASs use a variety of linguistic tools that be able to deal with small fragments of text. Therefore, to retrieve the documents which contains the answer from a large document collections, QASs employ Information Retrieval (IR) techniques to minimize the number of documents collections to a treatable amount of relevant text. In this paper, we propose a model for passage retrieval model that do this task with a better performance for the purpose of Arabic QASs. We first segment each the top five ranked documents returned by the IR module into passages. Then, we compute the similarity score between the user's question terms and each passage. The top five passages (with high similarity score) are retrieved are retrieved. Finally, Answer Extraction techniques are applied to extract the final answer. Our method achieved an average for precision of 87.25%, Recall of 86.2% and F1-measure of 87%.**

***Index Terms*—Answer Extraction ,Information Retrieval, Passage Retrieval, Question Answering Systems.**

## I. INTRODUCTION

For the purpose of satisfying user requirements, a user submit a query to the Information Retrieval (IR) systems, and the IR system has to return a set of documents arranged by their connectivity to the user's query. Several techniques are used to perform the process of document extraction, but most of majority of them depends on pattern matching technique that is based on the number of times that a user's query terms occur in each document, in addition to significance or differentiation value of these terms in the document collection. On the other hand, Question Answering systems (QASs) try to improve the output produces by IR systems by only retrieving small pieces of text that are likely to contain the information that satisfy the question/query. In fact, QASs consists of IR and Natural Language Processing(NLP) techniques to carry out their task. This combination of IR & NLP allows the minimization of the answer to the lowest level that allows accurate answer detection and extraction. However, knowing the fact that NLP techniques are computationally expensive, QASs need to minimize the amount of text in times when these techniques need to be applied. Hence, they normally applied on the IR systems output [1] that return the most relevant documents to the user's query considering that they will contain the required. Most used IR systems are basically depending on three models: the cosine model, the pivoted cosine model, and the probabilistic model [2] [3] [4]. In addition, IR systems usually apply query expansion techniques that increase their precision. These techniques may be based on thesaurus [5] or on the combination of the most occurring terms in the top N relevant documents [6]. In the literature, there several Passage Retrieval (PR) systems that have also been proposed for English language[7][8]. PR systems works on portions of text so that they can determine the relevance of a document to a query, besides detecting document fragments that are likely to contain the required answer(rather than the full document). Although PR systems use IR-based methods to do their work, they are considered to be more effective than IR systems for QA tasks. In this paper, we are proposing a model for PR system for Arabic QA. The following section briefly discuss the backgrounds in IR, PR and QA. Section 3 shows the architecture of IR-n. Section 4 presents the evaluation accomplished and finally, section 5 details conclusions and work in progress. Open domain QASs are defined as software capable of generating the answer to user questions directly from open domain documents. In other words, systems that can extract small fragments from texts, in a way it is possible to infer the answer to a specific question from whose content of these fragments. As a benefit, these systems attempts to reduce the amount of time to find a concrete information. QASs differs from the usual search engines like Google and Yahoo. Such search engines returns a list of links in response to a user query and it is the sole responsibility of the user to go through these links and find the expected answer. In contrary to these traditional search engines , QA systems tries to find a precise answer to the user query. For a QA system to successfully satisfy a user's request, it is required by the QA system to understand questions completely. From a linguistic point of view, "understanding" means to perform several and typical steps on processing of natural language : lexical, syntactic and semantic. This processing takes much more time than the statistical analysis that is performed in IR. In addition, as QA systems have to deal with much text as done for IR tasks, and the user requires the answer in a short period of time, it is obligatory that first, an IR system process the query and then, the QA system process deals with its output. In this procedure, the time needed for analysis is highly decreased[9] .

***Waheeb Ahmed**, Department of Information Technology, Kannur University., Kannur, Kerala, India, 7994702655,*
***Dr. Babu Anto P**, Department of Information Technology, Kannur University, Kannur, Kerala, India, 9447405362.*

## II. METHODOLOGY

Our QA system consists of the following modules:

- Question Analysis.
- Document Retrieval.
- Passage Retrieval/Selection.
- Answer Extraction.

### A. Question analysis

Question analysis module receive natural language questions and converts them into queries for the document retriever. This process includes mapping the user's question into query by employing sophisticated question analysis techniques to generate complex and structured queries. As a part of its function, this module also identifies the expected answer type of a question. For example, the expected answer type of "ما هي الأخلاقيات؟"("What is ethology?") is DESCRIPTION:defintion. That is, the question is requiring a descriptive answer of type "definition". This information assists in guiding the answer extraction module to apply the proper technique for extracting the answer. Another example,"ما هي الولاية الأمريكية التي عاشت بستة أعلام؟" ( "What U.S. state lived under six flags ?"). The expected answer type for this question is "LOCATION:state". This indicates that the question is looking for a location, specifically a "state" name.

### B. Document retrieval

This module returns a list of ranked relevant documents from the collection of documents /corpus. This process reduces the corpus to a manageable set of documents for additional processing. We used Vector Space Model(VSM) [2]for document retrieval for its simplicity and efficiency.

### C. Passage retrieval

This module analyze the retrieved set of documents and returns a list of ranked passages scored with respect to query terms. This module extracts the passages based on its similarity with the user's question. This module works basically by document segmentation. For segmentation of documents , our method first identifies sentences in the document using punctuation marks as separators, and then it identifies paragraphs by means of separators made by empty lines. The passages are detected by merging paragraph boundaries into segments until they satisfy the required length ,e.g..500 characters). The separate passages do not have any common content with each other, and the sliding passages change in size with the difference of one paragraph boundary, i.e., we start creating a new passage starting from the initial beginning of each paragraph of the document. If the boundaries of these paragraph are not detected, then these adapting(sliding) passages are getting partially overlapped with each other.

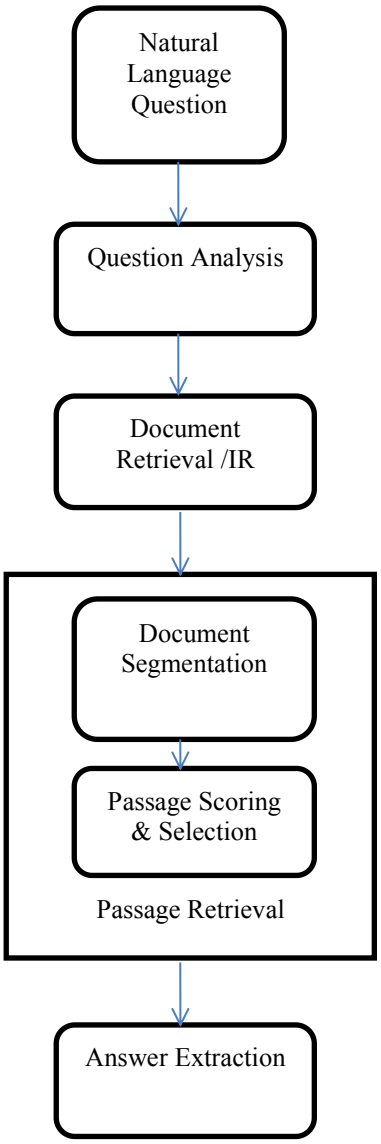


Fig. 1 QA system architecture with Passage Retrieval

The technique for passage retrieval works as follows:

1) Question terms are arranged based on the number of documents they occur in. Terms that appear in fewer documents are processed firstly.
2) The documents which contain any term of the question are selected.
3) The similarity for each passage p (available in the retrieved/selected documents) with the question q is calculated using the following formulas:
   Tf-Idf= tf x idf (1)
   $tf(t,d) = 1 + \log f_{t,d}$ (2)
   $idf=\log(1+N/n_t)$ (3)
   Where,
   Term frequency(tf)=the no. of times a term appears in a document
   inverse-document frequency(idf)=the no. of documents in which the term appears Only the most relevant passage of each document is selected for retrieval. That is, the most similar passage to the question terms is returned.
4) The passages that are selected are sorted by their similarity measure. The passage with the highest

similarity score is put at the top of the list in descending order.

5) Passages are connected with the document they belongs to and they are sorted in a ranked list.
6) We used the similarity measure (the cosine measure) presented in [2]. With one difference is that the size of each passage (the count of terms) is not used to normalize the results. In this proposed method we are using N as the number of documents in the collection, instead of the number of passages based on the methodology proposed in [9].

### D. Answer extraction

Answer Extraction module process the top ranked passages for extracting the final answer to the user's question. Typically, named-entity recognition technique is used to find candidate answers that match the question's named entities. For questions that are looking for dates a pattern matching technique is used. Questions that requires descriptive answers like how & why cosine similarity is used to find the answers.

## III. Results and Evaluation

Our method is tested using 200 questions and the corpus consisted of 500 documents extracted from Arabic Wikipedia. The measures used for calculating the performance of our passage retrieval technique are the same which used for evaluating TREC-10 [10] project:

Precision is a measure of the capability of the system to return only relevant passages.

$$\text{Precision}=\frac{No.of\ relevant\ passages\ retrieved}{Total\ number\ of\ Passages\ retrieved}$$

$$\text{Recall}=\frac{No.of\ relevant\ passages\ retrieved}{No.of\ relevant\ passages\ in\ collection}$$

$$\text{F1-measure}=2x\ \frac{Precision\ x\ Recall}{Precision+Recall}$$

Table 1 The precision, recall and F1 measure after using the passage retrieval technique.

| Question set | Precision(%) | Recall(%) | F1-measure |
|---|---|---|---|
| First set (50 questions) | 84 | 79 | 81.4 |
| Second set (50 questions) | 86 | 92 | 88.8 |
| Third set (50 questions) | 91 | 85 | 87.9 |
| Fourth set (50 questions) | 88 | 89 | 88.5 |
| **AVERAGE** | **87.25** | **86.2** | **87** |

Table 1 shows the different sets of questions along with their scored values by the QA system for precision, recall, and F1-measure. At bottom of the table is the average value for each one of them which is considered the average final performance of our QA system which employs both the passage retrieval method and answer extraction techniques.

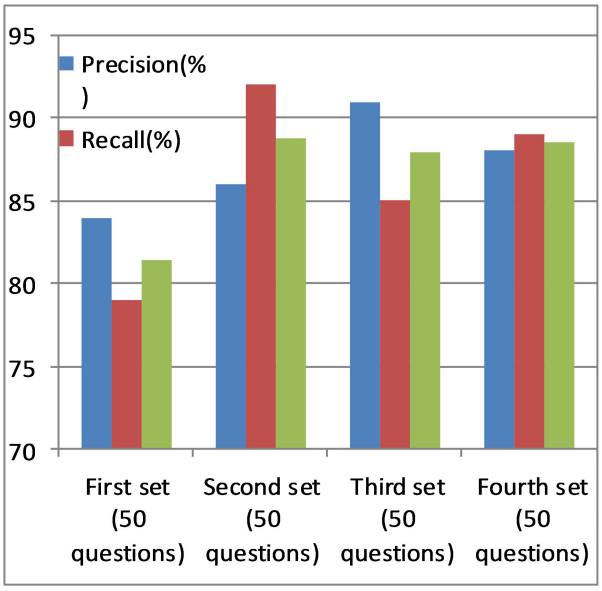


Fig. 2 Performance distribution of our QA system employing passage retrieval technique.

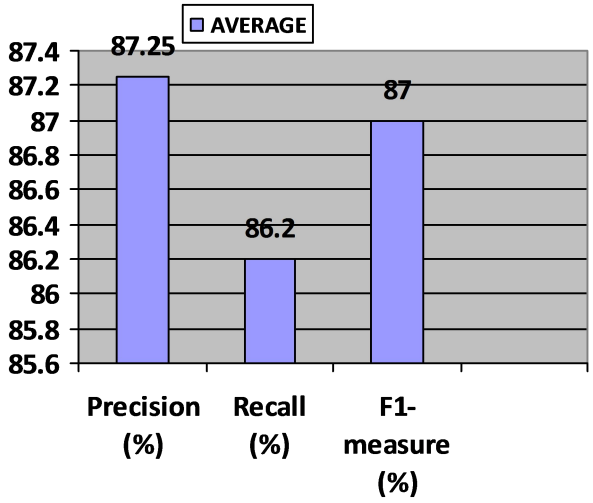


Fig. 3. Distribution of the average value for precision, recall and F1-measure.

Fig. 2 shows the performance of our QA systems after submitting 200 questions in four separate runs. The result is generally promising comparing to existing QA systems in different languages like English.

Fig. 3. Shows the average values of precision, recall and f1-measure which is very promising result and can be used in the enhancement of Arabic question Answering systems.

## IV. Conclusion

The obtained results in this paper shows that it is possible to highly improve answer passage retrieval and ranking by incorporating several factors including document segmentation into passages, measuring similarity of the passages with the user question and ranking the candidate answers. Our passage retrieval technique can be used for developing efficient Arabic QA

systems.

## REFERENCES

[1] A. Lally, J. M. Prager, M. C. McCord, B. K. Boguraev, S. Patwardhan, J. Fan P. Fodor and J. Chu-Carroll. ," Question analysis: How Watson reads a clue", IBM Journal of Research & Development, Vol. 56 ,No. 2, July 2012.

[2] G Salton. "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer" ,Addison Wesley Publishing, New York. 1989.

[3] A. Singhal, C. Buckley and M. Mitra, "Pivoted document length normalization.", Proceedings of the 19th annual international ACM- 1996.

[4] G. Venner, S. Walker and S. Okapi ,"Best match System". Microcomputer networking in libraries II. Vine, 48, pp. 22-26.,

[5] 1983.

[6] W. Mohammed, A. Basim and A. Adnan, "The effect of using a thesaurus in Arabic information retrieval system", International Journal of Computer Science Issues(IJCSI), Vol. 9, No. 1, pp. 431-435., November 2012.

[7] R. Cummins, & C. O'Riordan, "An axiomatic comparison of learned term-weighting schemes in information retrieval: clarifications and extensions". Artificial Intelligence. Rev., Vol. 28, No. 1, pp. 51–68, 2007.

[8] A. Mahboob and V. Suzan, "Passage Retrieval for Question Answering using Sliding Windows", In Proceedings of the 2nd workshop on Information Retrieval for Question Answering(Coling 2008), pp. 26-33, August, 2008.

[9] Hong Sun et. al. ," Answer Extraction from Passage Graph for Question Answering", In Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, pp. 2169-2175, August 2013.

[10] W. Bruce Croft, "Answer models for question answering passage retrieval", Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 516-517, July 2004.

[11] E. Voorhees," Overview of the TREC 2001 question answering track," Proceedings of the 10th Text Retrieval Conference, pp. 42–52, 2001.