Volume 3, Issue 4, April 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Performance Analysis of Effective Arabic Language Question Answering System

Jaspreet Kaur[#1], Vishal Gupta[*2]
[#]*ME Student, Dept. of Computer Science and Engineering, Panjab University Chandigarh, India*
[*]*Assistant Professor, Dept. of Computer Science and Engineering, Panjab University Chandigarh, India*

Abstract — With technological development, Question Answering has come out as the major area for the researchers. In Question Answering user is provided with specific answers instead of large number of documents or passages. Question Answering offers the solution to get effective and accurate answers to user question asked in natural language rather than language query. Arabic is among the languages that are less concerned by researchers in area of Question Answering System This paper compares some already existing Question Answering Systems for Arabic language. Different Arabic language Question Answering Systems are compared on performance basis. We also talk about the best approach out of all and the reasons for its effective performance and some methods for additional improvement in those techniques.

Keywords — Arabic, Question-Answering, Performance, Algorithms.

I. INTRODUCTION

In this paper, we examine the previous works on Arabic Language Question Answering systems and assess their performances and suggest different measures for further improvement. Comparing Arabic Question Answering Systems with each other, Arabic language is a key problem. We review studies on different techniques used for Arabic Question Answering System and discuss important issues which are helpful for building QA systems. Very less amount of work has been done in this language and thus needs improvement in lots of areas. Following are the Question Ansering Systems that we studied and performed comparison analysis.

II. Experimenting With A Question Answering System For The Arabic Language

An approach to building a question answering system is described called QARAB that provides short answers to questions expressed in the Arabic language. The system utilizes techniques from Information Retrieval and Natural Language Processing to process a collection of Arabic text documents as its primary source of knowledge. Initial results and analysis seem to be promising. Evaluation process of the system was based on 113 questions and a set of documents collected from newspaper AL-Raya. Average length of the answers obtained was 31 words.

A. QARAB Architecture

The system accepts natural language questions expressed in Arabic language and tries to find short text passages that answer the questions. The important modules of the QARAB Question Answering system are shown in Figure 1.
The key task can be summarized as follows: Given a group of questions in Arabic, identify short answers to the questions under the following two assumptions [5]:

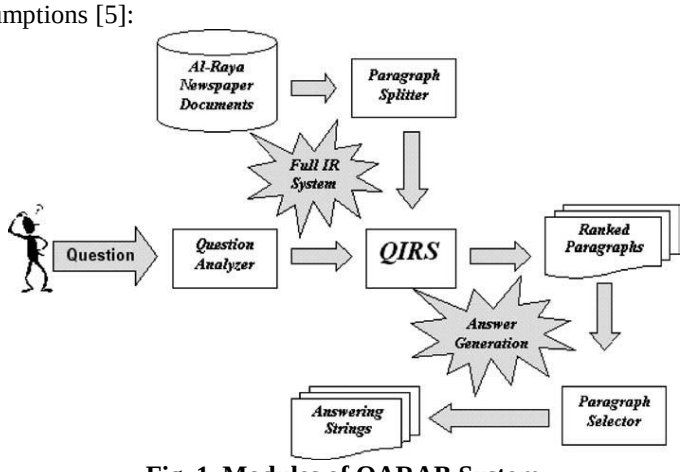

Fig. 1. Modules of QARAB System

A1. The answer string is present in a collection of Arabic newspaper articles extracted from the Al-Raya newspaper published in Qatar.
A2. The answer does not extend over several documents.
The fundamental QA process in QARAB is composed of four major steps:
S1. Processing the input question
S2. Matching the key elements of the question against the database of passages
S3. Recovering the candidate passages which contain potential answers using the IR system
S4. Processing each candidate passage in the same way as the question is processed and delivering a ranked list of five < answer-string, document_id> pairs such that each answer string is considered to contain an answer to the question and the identifier of the document supporting that answer.
The processing of the input question and candidate passages retrieved by the Information Retrieval system is carried out using Abuleil's tagger [1]. The tagger tags each word with part-of-speech information and adds it to the lexicon if it is not present. The tagger contains a huge table of keywords that it uses to recognize various types of proper nouns, so that it can identify new ones and construct lexical entries for them on the fly. Open questions include names of people, cities, countries, organizations, dates and events, the recognition of proper names. As Arabic text does not differentiate between upper and lower case letters, Abuleil's work [2] on recognizing proper nouns via keywords is an essential constituent of the success of QIRS module, a special-purpose QARAB Information Retrieval System, which was built from scratch to extract Arabic documents.

1) Question Analyzer
The question is tagged and parsed to decide its category and the type of answer it look for. Like other QA systems, the QARAB System employs standard Message Understanding Conference (MUC) categories [3]. The categories are: Person, Location, Organization, Percent, Date, Time, Duration, Measure, and Money. Questions in QARAB are classified based on the set of the question types given in Table 1. With help of question the type of processing needed to identify and extract the final answer can be determined. QARAB handles the incoming question as a ''bag of words'' against which the index file is searched to get a list of ranked passages that contains the answer. The question processing starts by performing tokenization to get individual terms. Effort is made on identifying proper names, as they are our best guides in identifying a possible answer. The query can be expanded to include all the words that exist in the word index and share the same roots as the query words to achieve better results.
The output of the query processing is sent to the QIRS system to get a ranked list of passages that match the terms of the query. The stemmer and the root-finder have been modified to give perfect results [5].

2) Arabic Information Retrieval System (QIRS)
For a successful QA system a good guide is needed to find the documents/passages that are most relevant to the question and thus, decrease the number of documents/passages to get an answer.

TABLE I
QUESTION TYPES PROCESSED BY QARAB SYSTEM

Question starting with	Question Type
Who, Whose	Person
When	Date, Time
Where	Location
How much, How many	Number, Quantity

The QIRS Information Retrieval system is implemented from scratch [4] and is based on Salton's vector space model [5]. The QIRS system uses a relational database management system, MSSQL 7, to store not the inverted index files and a number of temporary files that are too large to fit in memory. The QIRS system was built to experiment with a light-stemming algorithm, which covers the common suffixes/prefixes from a word to construct the modified-word index and a root based algorithm, which manages each word to extract its root to build the root index. QARAB used QIRS to extract the ten passages that best match the question bag of words ranking them using the familiar Inverse Document Frequency method [5].

3) Answer Generator
Now parsing is done on the ten candidate passages extracted by the QIRS system to identify named entities matching the category type expected by the query and check whether the sentences are qualified to be possible answers. After this filtering process, the first five answers that stays, are presented to the user as <answer-string, document-ID> pairs. Each one of the five strings is examined by the user to find the answer that exactly answers the question. Abuleil's tagger [1] is used in processing the query and the candidate answer passages retrieved by the IR system. The tagger works on the original text for the query and the paragraphs extracted by QIRS. The tagger moves through the text forming lexical entries for every new word it gets and adds attributes as it finds them. In Arabic since there is no difference between upper and lower case characters, that is why proper nouns can only be identified by associated key words [5].

B. Experimental Results

Two runs were carried out using QARAB system. The first run includes modified-word strategy (word index), while the second run was a query expansion using the root-based strategy (root index). The system was examined by four native speakers of Arabic who asked a set of 113 questions to the system.

1) Experiment 1

In this experiment, the light stemmer was used to process 113 questions presented by native speakers of Arabic, all with a university education. The speakers were not specified the articles, but the time period covered by the articles was known. The question bag of words was provided to the QIRS system to retrieve the necessary paragraphs, which probably contain an answer to the question. The system then sorts out the paragraphs and, for each question, returns a ranked list of five <answer-string, document-id> pairs [5]. Answering strings and the supporting Doc-Id's were saved in a log file for every question. The system reports this failure in case no answer is found. At the end of the first run, the log file is analyzed to obtain the final judgment from the user's point of view.

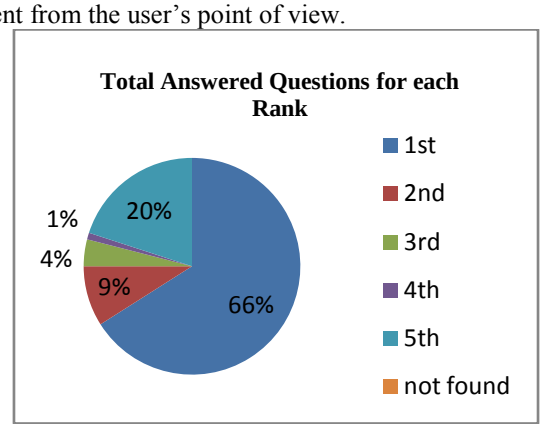

Fig. 2. Rank percentage of the Questions Answered in Experiment 1

90 questions out of the 113 questions in the test collection were answered accurately. 75 questions were answered by the first string returned, 10 questions by the second string, 4 questions by the third string, while just 1 question was answered by the fifth string. The system missed 23 questions as shown in Figure 2. The performance of the system was measured with respect to the two parameters: recall and precision. The definitions of recall and precision in the QA context were presented by Gaizauskas and Humphreys [6]. These measures are calculated using the formulas:

$$\text{recall}_{QA} = \frac{\text{number of correct answers}}{\text{number of questions to be answered}}$$

$$\text{precision}_{QA} = \frac{\text{number of correct answers}}{\text{number of question answered}}$$

Precision_{QA} measures whether the answers extracted are correct or not. The MRR is the standard TREC effectiveness measure for each TREC QA run [7]. Each question was assigned a score equal to the inverse of the rank of the first string that was judged to contain a correct answer in order to calculate MRR. The question was assigned a score of zero if none of the five answer strings contained an answer. The MRR value for the run is calculated by taking the average of the scores for all the questions. Now let us compute the recall_{QA} of the first run:

$$\text{recall}_{QA} = {}^{90}/_{113} = 79.6\%$$

Now precision_{QA} is calculated as follows:

$$\text{precision}_{QA} = {}^{90}/_{90} = 100\%$$

90 questions were answered correctly out of the 90 questions that were answered in the first run.
Table 2 shows the recall_{QA} percentage of experiment.

TABLE II
RECALL_{QA} PERCENTAGE USING THE QUESTION WORDS

Question Type	Total Questions	Number Of Correct answers	Recall_{QA}	Average Length of the Questions	Average Length of the Answers (words)
Who	58	56	96.5%	4	34
When	18	10	55.6%	4	37
Where	14	12	85.7%	4	31
How much/many	23	12	52.2%	8	28
Overall Results	113	90	79.6%	5	33

System failed to to answer 23 questions due to variations in verb tenses or word inflections between the question bag of words and the words appearing in the supporting documents. The $precision_{QA}$ of the first run was 100% and MRR was 0.718 as shown in Table 3.

TABLE III
$PRECISION_{QA}$ PERCENTAGE AND MEAN RECIPROCAL RANK (MRR) IN EXPERIMENT 1

Question Type	MRR	Number of correct answers	Number of questions answered	$Precision_{QA}$
Who	0.848	56	56	100%
When	0.50	10	10	100%
Where	0.782	12	12	100%
How much/many	0.522	12	12	100%
Overall Results	0.718	90	90	100%

2) Experiment 2

The aim of this experiment was to elaborate the search by retrieving the roots of the words in the original question then adding to the query all the words from the word index that share the same roots as the words in the query. 110 questions out of the 113 questions were answered correctly. 90 questions were answered from the first string, 8 questions from the second string, 5 questions from the third string, 3 questions from the fourth string and 4 questions were answered from the fifth string. 3 questions were missed as shown in Figure 3. Now let us calculate the $recall_{QA}$ and $precision_{QA}$ of the second run:

$$recall_{QA} = {}^{110}/_{113} = 97.3\%$$

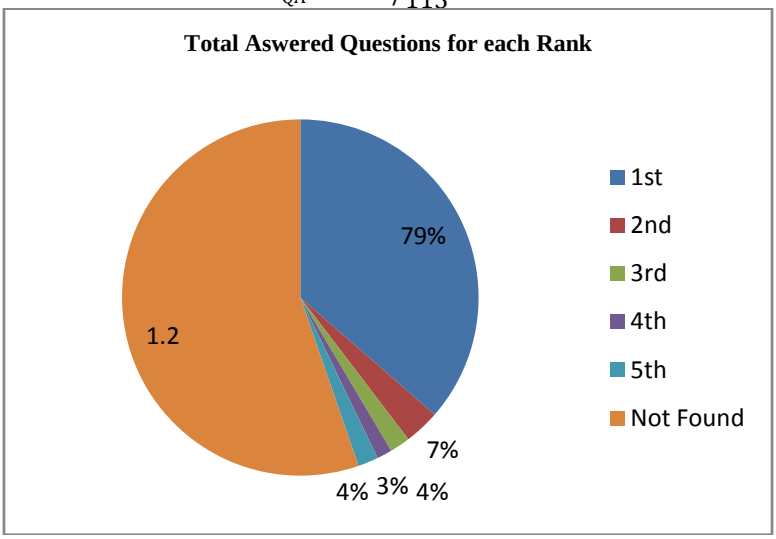

Fig. 3. Rank Percentage of the Questions Answered in Experiment 2

Table 4 shows the $recall_{QA}$ percentage of Experiment II and the $recall_{QA}$ percentage for each question type. Now let us calculate the $precision_{QA}$ of the second run:

$$precision_{QA} = {}^{110}/_{113} = 97.3\%$$

TABLE IV
$RECALL_{QA}$ PERCENTAGE IN EXPERIMENT 2 USING QUESTION EXPANSION

Question Type	Total Questions	Number Of Correct answers	$Recall_{QA}$	Average Length of the Questions	Average Length of the Answers
Who	58	55	94.8%	4	33
When	18	18	100%	4	31
Where	14	14	100%	4	34
How much/many	23	23	100%	8	30
Overall Results	113	110	97.3%	5	31

TABLE V
PRECSION$_{QA}$ PERCENTAGE AND MRR IN EXPERIMENT 2 USING QUESTION EXPANSION

Question Type	MRR	Number of correct answers	Number of questions answered	Precision$_{QA}$
Who	0.783	55	58	94.8%
When	0.863	18	18	100%
Where	0.946	14	14	100%
How much/many	1	23	23	100%
Overall Results	0.860	110	113	97.3%

110 questions out of the 113 questions were answered correctly [5]. 3 questions that were missed in the second run did not have correct answers in any of the returned strings and therefore, they received a score of zero. The precision$_{QA}$ of the second run was 97.3% and MRR was 0.860 as shown in Table 5.

III. Defarabicqa: Arabic Definition Question Answering System

In this paper a definitional Question Answering system called DefArabicQA is proposed. This system presents effective and accurate answers to definition questions expressed in Arabic language from Web resources. It is based on an approach that uses a little linguistic analysis and no language understanding capability. DefArabicQA finds candidate definitions by using a set of lexical patterns and categorize these candidate definitions by using heuristic rules and ranks them by using a statistical approach. Two experiments have been carried out on DefArabicQA [7]. Experiment 1 was based on Google as a Web resource and has obtained an MRR equal to 0.70. Experiment 2 was based on Google coupled with Wikipedia as Web resources and has obtained MRR equal to 0.81. Improvement in the quality of the definitions needs to be done because experimental evaluation is conducted on the basis of questions asked by the native speaker. In some cases, few words are missed at the end of the definition answer. This is because of the fact that the snippet itself is truncated. An empirical study is needed to determine different weights to the three used criteria for ranking the candidate definitions.

A. The DefArabicQA System

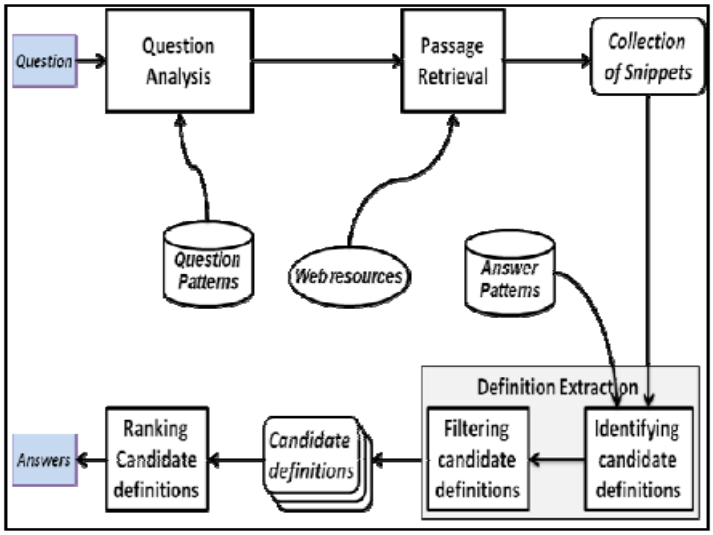

Fig. 4. Architecture of DefArabicQA

The basic architecture of the DefArabicQA system is shown in Figure 4. The system consists of the following modules: i) Question Analysis, ii) Passage Retrieval, iii) Definition Extraction and iv) Ranking Candidate definitions [4] [8].

1) Question Analysis: This module is a fundamental module of DefArabicQA. The purpose of this module is to recognize the topic question and decide the answer type expected. The question topic is recognized by using two lexical question patterns Table 6 and the answer type expected is inferred from the interrogative pronoun of the question.

2) Passage Retrieval: The passage retrieval module gets the top-n snippets extracted by the Web search engine. This specific query is consists of the question topic which is recognized by the question analysis module. After gathering the top-n snippets, only snippets having the integrate question topic are kept on the basis of some heuristic.

TABLE VI
QUESTION PATTERNS AND THEIR EXPECTED ANSWER TYPES USED BY DEFARABICQA SYSTEM

Question Patterns	**Expected Answer Type**
Who+be+<topic>?	Person
What+be+<topic>?	Organization

3) Passage Retrieval: The passage retrieval module gets the top-n snippets extracted by the Web search engine. This specific query is consists of the question topic which is recognized by the question analysis module. After gathering the top-n snippets, only snippets having the integrate question topic are kept on the basis of some heuristic.

4) Definition Extraction: This module is the heart of a defArabicQA system and it consists of two sub-modules: i) identifying candidate definitions, and ii) filtering candidate definitions.

*3.1) Identifying Candidate Definitions:*In this part, we recognize and extract candidate definitions from the group of snippets collected in the passage retrieval module. Lexical patterns are used to recognize these candidate definitions. Basically, a lexical pattern is a series of strings that give a context to recognize the correct answers. It reveals a common use of written ways used to pioneer an organization. Patterns are formed manually and no natural language processing is used in their construction. Candidate definition is recognized by a particular pattern if the surrounding of the question topic in a snippet is known by a specific pattern.

3.2) Filtering Candidate Definitions: Heuristic rules are employed to filter the identified candidate definitions. These heuristic rules are inferred from the examination of a group of annotated candidate definitions.

4) Definition Ranking: Definition Ranking is built on an algebraic approach. A global score is employed to rank candidate definitions retained in the Definition Extraction module. Global score is a combination of three scores related to three measures of a candidate definition: i) pattern weight criterion, ii) snippet position criterion, and iii) word frequency criterion. First top-5 candidate definitions ranked according to their global scores are presented to the user.

4.1) Pattern weight criterion (C_1): The score of this measure is the weight of the pattern that has identified the candidate definition CD_i. This score is denoted by:

$$C_1 (CD_i) = w_i \quad (1)$$

Where w_i, presents the weight of pattern i. A weight is associated to each pattern according to its significance.

4.2) Snippets Position Criterion (C_2) : The score of this measure represents the position of the snippet that is having the candidate definition in the snippets collection. This score denoted by:

$$C_2 (CD_i) = p_i \quad (2)$$

Where p_i, is the position of the snippet that contains the candidate definition CD_i.

4.3) Word Frequency Criterion (C_3): The score of this measure represents the sum of the frequencies of the words that occur in a candidate definition. Firstly, a centered vector containing common words across candidate definitions with their frequencies is constructed, beyond stop words. Secondly, the frequency sum of the words recurring in both CD_i and centered vector is calculated. The candidate definition CD_i score is calculated as given below:

$$C_3 (CD_i) = \sum_{k-1}^{n} f_{ik} \quad (3)$$

Where n is the number of words occurring in centered vector and in CD_i, the candidate definition, $1 \leq k \leq n$ and f_{ik} is the frequency of word k.

4.4) Criterion Aggregation: To facilitate the three measures explained above, firstly the normalization of the score of each measure is divided by the maximum score as follows:

$$C'_{i,j} = {C_{i,j}}/{MaxC_i} \quad (4)$$

Where i is a candidate definition and j is a criterion.
Secondly, the three normalized scores are combined to get the global score GS of the candidate definition CD_i.

$$GS(CD_i) = \sum_{j=1}^{3} C'_{i,j} \quad (5)$$

B. Experimental Results

Two experiments were carried out using the *DefArabicQA* system. Experiment 1 was carried out using Google Search engine6, while the experiment 2 was carried out using Google Search engine and the free encyclopedia Wikipedia Arabic version7. 50 organization definition questions were used. An Arabic native speaker accessed the system. MRR is used as evaluation metrics. It is calculated as follows: each question is assigned a score equal to the inverse rank of the first string that is judged to contain a correct answer. If none of the five answer strings contain an answer, the question is assigned a score of zero. The MRR value for the experiment is calculated by taking the average of scores for all the questions.

1) Experiment 1 and its Results

From 50 questions in the test collection [7], 41 questions (82%) were answered correctly by complete definitions in the top-five candidate definitions. 27 questions were answered by the first candidate definition returned, 7 by the second candidate definition, 3 by the third candidate definition, 3 by the fourth candidate definition and 1 by the fifth candidate definition as shown in Table 7. The systems missed 18% of the questions as shown in Table 8. MRR was equal to 0.70 as shown in Table 9.

TABLE VII
ANSWERED QUESTIONS RATE FOR EACH RANK

	Experiment 1	**Experiment 2**
1st rank	27 (54%)	32 (64%)
2nd rank	7 (14%)	8 (16%)
3rd rank	3 (6%)	2 (4%)
4th rank	3 (6%)	1 (2%)
5th rank	1 (2%)	2 (4%)
Top-5	41 (82%)	45 (90%)

TABLE VIII
UNANSWERED QUESTIONS RATE

	Experiment 1	**Experiment 2**
Top-5	9 (18%)	5 (10%)

TABLE IX
MRR VALUES FOR BOTH EXPERIMENTS

	Experiment 1	**Experiment 2**
MRR	0.70	0.81

2) Experiment 2 and its Results

The main purpose of the second experiment is to calculate the significance added by the Web resource Wikipedia to the results obtained in the first experiment with the Google search engine [7]. Out of the 50 questions in the test collection, 45 questions were answered correctly by complete definitions in the top-five candidate definitions. 32 questions were answered by the first returned candidate definition, 8 by the second candidate definition, 2 by the third candidate definition, 1 by the fourth candidate definition and 2 by the fifth candidate definition as shown in Table 2. The system missed 5 questions as shown in Table 3. The value obtained for MRR is 0.81.

IV. Idraaq: New Arabic Question Answering System Based On Query Expansion And Passage Retrieval

A. Method

The word "IDRAAQ" in Arabic has the following meanings and senses: to understand, to recognize, to reach an objective, knowledge, intelligence, etc. The paper presents core modules of a new Arabic Question Answering system called IDRAAQ. These modules aim at enhancing the quality of retrieved passages with respect to a given question. Experiments have been conducted in the framework of the main task of QA4MRE@CLEF 2012 that includes this year the Arabic language. Two runs were submitted. Both runs only use reading test documents to answer questions. The difference between the two runs exists in the answer validation process which is more relaxed in the second run. The Passage Retrieval (PR) module of our system presents multi-levels of processing in order to improve the quality of returned passage and thereafter the performances of the whole system. The PR module of IDRAAQ is based on keyword-based and structure-based levels that respectively consist in: (i) a Query Expansion (QE) process relying on Arabic WordNet semantic relations; (ii) a Distance Density N-gram Model based passage retrieval system. The latter level uses passages retrieved on the basis of QE queries and re-ranks them according to a structure-based similarity score.

B. System Architecture

The IDRAAQ system is entirely programmed in Java. Other third party components and resources are also used. Three typical modules of a Question Answering system are designed as shown in Figure 5 [14].

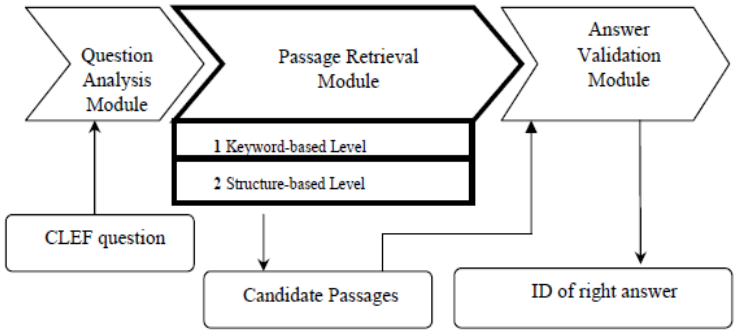

Fig. 5. Basic Architecture of IDRAAQ

1. Question Analysis and Classification module: Question is examined in order to obtain its keywords, recognize the structure of the expected answer and create the query to be passed to the passage retrieval module.

2. Passage Retrieval (PR) module: This module uses the query generated by the earlier module and obtains a list of paragraphs from an Information Retrieval process. After that ranking process is performed in order to increase the importance of the candidate passages according to the user question. The performance of the IDRAAQ system is mainly dependant on this module and on the quality of its returned passages. PR module of IDRAAQ is formed by two implemented levels: Keyword-based level and Structure-based level.

2.1 Keyword Based Level: This level is related to a semantic Query Expansion (QE) process. Every question keyword is replaced by its grammatically related terms that are retrieved from the Arabic Word Net (AWN) [10]. Four relations are used in this level in AWN: synonymy, hyponymy, hypernymy and SUMO-AWN relations. SUMO (Suggested Upper Merged Ontology) is a high level ontology planned with AWN synsets4. The objective of this QE process [14] is illustrated in Figure 6.

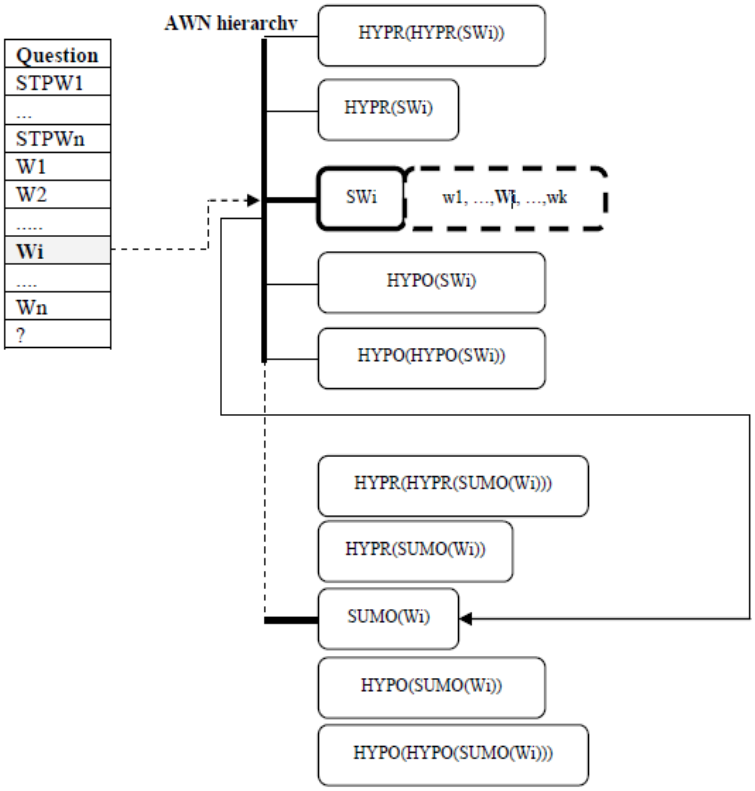

Fig. 6. AWN-based QE integrated in the IDRAAQ System

From each question, only non stopwords (STPWn) in the QE process are taken into consideration. AWN-based QE process takes as input an Arabic word, Wi and generates the following terms:
a. Morphological variants of Wi.
b. Words that share the same AWN synsets (SWi) with Wi (the synonyms w1…wk).
c. Words that share the AWN synsets that are hyponyms of each SWi denoted by HYPO (SWi).
d. Words that share the AWN synsets that are hypernyms of each SWi are referred to by HYPR (SWi).
e. Words that appear in the definition of the SUMO concept which is equivalent to each SWi.
The similar process is repeated for words associated to HYPO(SWi) and HYPR(SWi). Just move 2 levels up and down in the AWN hierarchy to avoid endless recursive process. A list of words is generated for every question keyword that represents the perspective of the keyword in the AWN hierarchy. Words belonging to the context of the expanded word are retrieved by moving up and down in the AWN hierarchy. Other semantically related words can be extracted by moving to the synsets that are equivalent to these latter concepts. In case of Named Entities (NEs) keywords, the keyword is just replaced by its synonyms. The hypernyms are just inserted before the keyword in the question. IDRAAQ uses an augmented version of AWN. This augmentation mainly concerns NEs, noun hyponymy relations and verbs.

2.2 Structure Based Level: For each question, distinct new queries are formed according to the terms retrieved from AWN. The structure-based level introduces a new measure to effectively re rank passages: the Distance N-gram Density [11]. This model take into account sequence of n-gram extracted from a sentence or a question. Every feasible n-grams of the question are examined. A score is assigned according to the n-grams and weight that appear in the extracted paragraphs. If a paragraph consists of one or more related terms then it is extracted. In the IDRAAQ system, this model is

implemented by means of the Java Information [12]. The JIRS is incorporated in the PR module of IDRAAQ by following steps as given below:

Step 1: Retrieve associated queries of a question.
Step 2: The list of queries is designed using the JIRS input file.
Step 3: Documents are also configured using the SGML JIRS format.
Step 4: The list built in step 3 is cataloged using the related JIRS process.
Step 5: The JIRS process is implemented on the cataloged collection and using the input file.
Step 6: Out of all the queries, the five passages, with the best JIRS similarity score, are taken into account in the Answer Validation module.

3. Answer Validation (AV) module: It validates an answer from a list of candidate answers depending on passages that are generated by the previous module. The performance of the IDRAAQ system mainly depends on PR module and on the quality of its returned paragraphs.

C. Experimental Evaluation and Results

Test set at 2012 consists of 4 topics; every topic consist 4 reading tests. Every reading test contains one document, followed by 10 questions, each with a group of 5 answer choices per question. There are 16 test documents having 4 documents for all four topics, 160 questions having 10 questions for every document and 800 options i.e. 5 for every question for each language task. Questions have the following characteristics:

a. Multiple choice questions are there; where each question has 5 possible answers.
b. They are designed in such a way so that it can focus on testing the comprehension of one single document.
c. Test the reasoning capabilities of systems.

Questions may be of the following types:

a. Factoid: Where or When or By--Whom
b. Causal: What was the cause of Event A?
c. Method: How did A do B?
d. Purpose: Why was A brought about?
e. Which is true: Here one must select the correct option from a number of statements.

The IDRAAQ system applies for each question the preprocessing stage, the keyword based stage and the structure-based stage. The answer checking process matches candidate answers with returned passages. Each test gets an evaluation score between 0 and 1 using c@1 [13]. Systems receive evaluation scores from two different perspectives: a). at the question-answering level: correct answers are counted individually without grouping them and b). At the reading-test level: figures both for each reading test as a whole and for each separate topic are given.

Two measures have been evaluated as follows:

1. Overall Accuracy:

$$\text{Accuracy} = \text{nr}/\text{n}$$

Where nr is the number of correctly answered questions and n is the total number of questions.

2. The c@1 measure:

$$\text{C@1} = (\text{nr} + \text{nu} * \left(\frac{\text{nr}}{\text{n}}\right))/\text{n}$$

where nu is the number of unanswered questions.

Obtained results also presents number of unanswered question with right and wrong candidate answers. However, in both runs, we did not consider this possibility in the submitted outputs. Table 10 and 11 presents the obtained results in terms of: (i) accuracy over all questions and (ii) the overall as well as detailed c@1 measure [14].

TABLE X
OVERALL ACCURACY OF IDRAAQ OVER THE TWO SUBMITTED RUNS

Runs	**Overall Accuracy**	**Answered**		**Unanswered**		
		Right	**Wrong**	**Empty**	**Right**	**Wrong**
Run1	0.08	12	21	127	0	0
Run2	0.13	21	49	90	0	0

TABLE XI
C@1 RELATED TO IDRAAQ

Runs	**c@1 Measure**				
	Overall	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**
Run1	0.13	0.25	0.18	0.05	0.05
Run2	0.21	0.36	0.19	0.08	0.17

As shown in Table 10 above, the overall accuracy achieved is 0.13 in the second run. This accuracy is computed over the 160 questions [14]. Table 2 shows the overall of 0.21 as of the second run (versus 0.13 for the first run). Figure 7 demonstrates a comparison between the best c@1 measures obtained over the four topics relating to this level. Topic #3 is the one for which lower performances have been reached. Most of the answered questions are factoid ones (When,

Who, What, etc.). This implies that using Arabic Word Net mapped with YAGO has a positive influence on system performances.

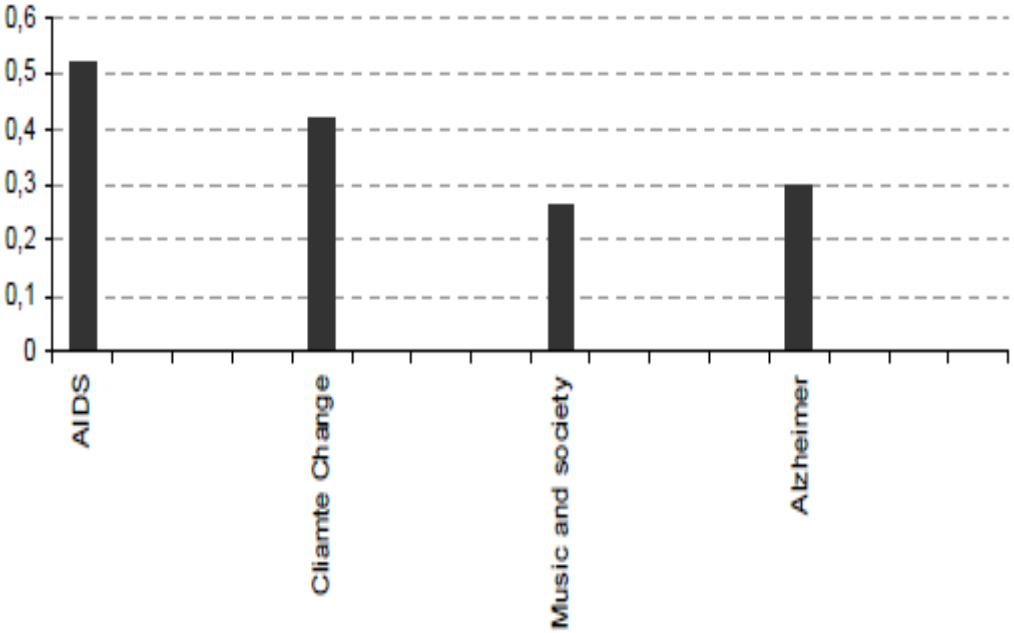

Fig. 7. Best c@1 Obtained in Reading Tests over Topics

V. **Performance Analysis Of The Above Mentioned Arabic Question Answering Systems**

Performance of different Arabic Question answering systems is shown in Figure 8. As we can see among three question answering systems QARAB is having the highest value of MRR but we consider IDRAAQ as the best of all although it has very low values of evaluation measures. This is because of the fact that the other two question answering systems are evaluated by native speaker and less number of questions is taken into consideration but IDRAAQ for the first time in Arabic language history has taken part in CLEF 2012. The results obtained for IDRAAQ at QA4MRE task of CLEF 2012 are encouraging. QARAB describes an approach to building a question answering system called QARAB which outputs short answers to questions expressed in the Arabic language. Two runs were carried out using QARAB system. The first run includes modified-word strategy (word index), while the second run was a query expansion using the root-based strategy (root index). First run obtained MRR of 0.718 and run 2 obtained value of MR equals to 0.86. Information Retrieval and Natural Language Processing techniques are utilized to process a collection of Arabic text documents as its primary source of knowledge. Initial results show potential. Users are all Arabic-speaking university graduates, who may be more relaxed with reading long answers and more anxious about the accuracy and the context of the material extracted than the typical user visualized by system designers. One of the issues that need to be considered is to experiment with elaborating the questions using a thesaurus based on Arabic lexical-semantic relations and measuring its effectiveness on the system. Number of question needs to be increased. Another possibility is to paraphrase the question collection set and measure the retrieval effectiveness of the system on the new set of questions. Improvement in quality of the answers is needed other question types, including what, why and how questions needs to be considered.

In DefArabic Question Answering System, an effective and accurate answers to definition questions expressed in Arabic language from Web resources. It is based on an approach that uses a little linguistic analysis and no language understanding capability. *DefArabicQA* finds candidate definitions by using a set of lexical patterns and categorize these candidate definitions by using heuristic rules and ranks them by using a statistical approach. Two experiments have been carried out on *DefArabicQA*. Experiment 1 was based on Google as a Web resource and has obtained an MRR equal to 0.70. Experiment 2 was based on Google coupled with Wikipedia as Web resources and has obtained MRR equal to 0.81. Improvement in the quality of the definitions needs to be done because experimental evaluation is conducted on the basis of questions asked by the native speaker. In some cases, few words are missed at the end of the definition answer. This is because of the fact that the snippet itself is truncated. An empirical study is needed to determine different weights to the three used criteria for ranking the candidate definitions.

The current edition of QA4MRE has considered for the first time the Arabic language and IDRAAQ has participated in this. In this semantic QE process is combined with the Distance N-gram Density model. The retrieved results are encouraging in particular for factoid questions. According to earlier preliminary experiments [15], the combination of the third level based on Conceptual Graphs and semantic similarity would enhance the performances of the system at the PR module as well as the Answer Validation module. The prospect of the current work is preparing the system in order to participate in the next edition of QA4MRE for Arabic in an aim of getting the best well-known QA systems for other languages. Two evaluation measures were used in this system that is accuracy and c@1 measure and obtained the values 0.13 and 0.21 respectively.

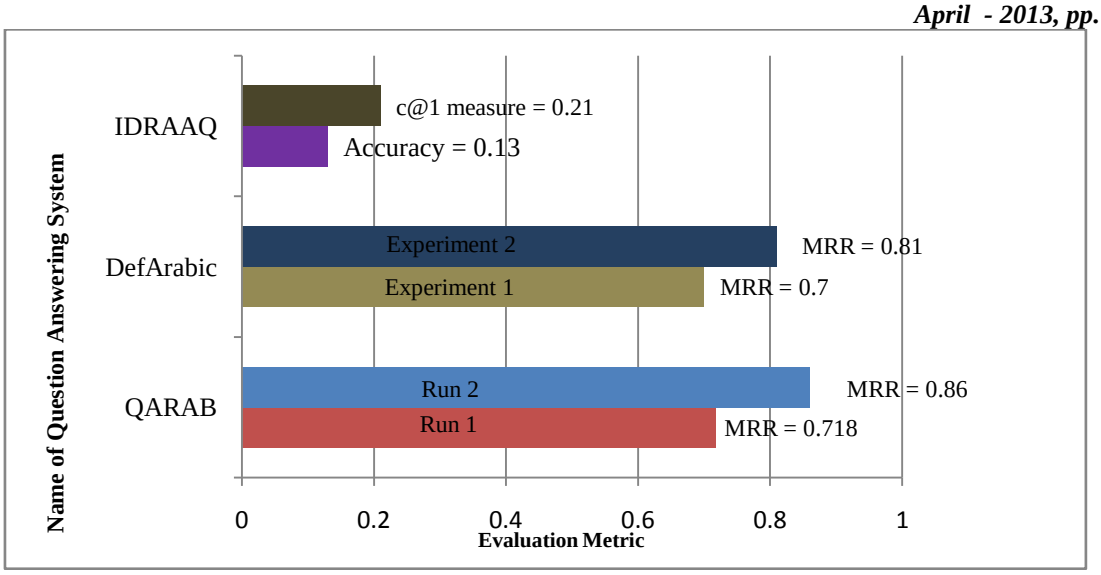

Fig. 8. Performance Analysis of Arabic Question Answering Systems

VI. Conclusion

We analysed that among the above described Arabic question answering systems, IDRAAQ (Information and Data Reasoning for Answering Arabic Questions) achieves accuracy and c@1 measure equals to 0.13 and 0.21 respectively. QARAB and DefArabic Question Answering Systems obtained higher values of MRR equals to 0.86 and 0.81 but we consider IDRAAQ as best among all Question Answering System as Arabic language was considered for CLEF 2012 for the first time and it achieved promising results at QA4MRE main task of CLEF 2012. By participating in the evaluation criterion, it realised its limitations and needs to work upon that in order to get better results in future. We consider QARAB and DefArabic QA next to IDRAAQ because both of systems performances were evaluated by the native speaker of Arabic language and less number of questions were taken into consideration. So, we conclude that very less amount of work has been done in Arabic Question Answering systems as compared to other languages and improvement in the systems being built in this language is being going on.

ACKNOWLEDGMENT

I would like to articulate my thanks to Mr. Vishal Gupta, Assistant Professor of Computer Science and Engineering Department in UIET, Department of Panjab University Chandigarh for his guidance in accomplishing this task.

REFERENCES

[1] S. Abuleil, and M. Evens, *"Extracting an Arabic Lexicon from Arabic Newspaper Text",* Computers and the Humanities, 36(3), pp. 191–221, 2002.

[2] S. Abuleil, K. Alsamara, and M. Evens, *"Tagging Proper Nouns and Keywords to Classify Arabic Newspaper Text",* in Proceedings of the 13th Midwest Artificial Intelligence and Cognitive Science Conference. Chicago, IL, pp. 137–142, 2002.

[3] N. Chinchor, "*Overview of MUC-7",* in Proceedings of the Seventh Message Understanding Conference, 1991.

[4] B. Hammo, H. Abu-Salem, S. Lytinen, and M. Evens, *"QARAB: A Question Answering System to Support the Arabic Language",* in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics: Workshop on Computational Approaches to Semitic Languages, ACL, Philadelphia, PA, pp. 55–65, 2002.

[5] B. Hammo, S. Abuleil, S. Lytinen, and M. Evens, *"Experimenting with a Question Answering System for the Arabic Language",* Computer and Humanities 38: 397-415, 2004.

[6] G. Salton, The SMART Retrieval System Experiments in Automatic Document Processing, Prentice Hall Inc., Englewood Cliffs, NJ, 1971.

[7] R. Gaizauskas, K. and Humphreys, *"A Combined IR/NLP Approach to Question Answering against Large Text Collections",* in Proceedings of RIAO, Content-Based Multimedia Information Access, Paris, France, April, pp. 1288–1304, 2000.

[8] E. Voorhees, *"Overview of the TREC 2001 Question Answering Track",* in Proceedings of the 10th Text REtrieval Conference (TREC 2001), NIST Special Publication 500–250, pp. 42–51, 2001. O. Trgui, L. H. Belguith, and P. Rosso, *"DefArabicQA: Arabic Definition Question Answering System",* in Proceedings of Workshop on Language Resources and Human Language Technologies for Semitic Languages, 7th LREC, Valletta, Malta, pp. 40—44, May 2010.

[9] Y. Benajiba, P. Rosso, and A. Lyhyaoui, *"Implementation of the ArabiQA Question Answering System's Components",* in Proceedings of Workshop on Arabic Natural Language Processing, 2nd Information Communication Technologies Int. Symposium, 2007.

[10] S. Elkateb, W. Black, P. Vossen, D. Farwell, H. Rodríguez, A. Pease, and M. Alkhalifa, *"Arabic WordNet and the Challenges of Arabic",* in proceedings of Arabic NLP/MT Conference, London, U.K, 2006.

[11] J. M. Gomez, M. Gomez, E. Sanchis, L. V. Pineda, and P. Rosso, *"Language independent passage retrieval for question answering",* in Fourth Mexican International Conference on Artificial IntelligenceMICAI, Lecture Notes in Computer Science, Springer Verlag, pages 816–823, Monterrey, Mexico, 2005.

[12] Y. Benajiba, P. Rosso and J. M. Gomez, *" Adapting JIRS Passage Retrieval System to the Arabic", i*n Proceeding of 8th International Conference on Computing Linguistics and Intelligent Text Processing, CICLing, Springer-Verlag, LNCS(4394), pp. 530-541, 2007.

[13] A. Penas, and A. Rodrigo, *"A Simple Measure to Assess Non-response",* in Proceedings of 49th Annual Meeting of the Association for Computational Linguistics- Human Language Technologies, Portland, Oregon, USA, June 19-24, 2011.

[14] A. Lahsen, K. Bouzoubaa, and P. Rosso, "IDRAAQ: New Arabic Question Answering Systen BASED ON Query Expansion and Passage Retrieval", in CLEF (Online Working Notes/Labs/Workshop), 17-20 September, Rome, Italy, 2012.

[15] L. Abouenour, K. Bouzoubaa, and P. Rosso, "*Structure-based evaluation of an Arabic semantic Query Expansion using the JIRS Passage Retrieval system*", in Proceedings of Workshop on Computational Approaches to Semantic Languages, Athens, Greece, April, 2009.

