Abstract

Today the Web is the largest resource of knowledge and, therefore, sometimes this makes it difficult to find precise information. Current search engines can only return ranked snippets containing the effective answers to a query user. But, they can not return the exact answers. Question Answering systems present the solution to obtain effective and exact answers to a user question asked in natural language question instead of keywords query. Unfortunately, Question Answering task for the Arabic language has not been investigated enough in the last decade, compared to other languages. In this paper, we tackle the definition Question Answering task for the Arabic language. We propose an Arabic definitional Question Answering system based on a pattern approach to identify exact and accurate definitions about organization using Web resources. We experimented this system using 2000 snippets returned by Google search engine and Wikipedia Arabic version and a set of 50 organization definition questions. The obtained results are very encouraging; (90%) of the questions used have complete (vital) definitions in the top-five answers and (64%) of them have complete definitions in the top-one answer. MRR was (0.81).

1 Introduction

Definition questions of the type ‘What is X?’ is frequently asked on the Web. This type of question is generally asked for information about organization or thing. Generally, dictionaries and encyclopaedias are the best resources for this type of answers. However, these resources often do not contain the last information about a specific organization or do not yet contain a definition of a new organization due to non instantaneous update. Thus, the user has the habit to look for a definition from searching the Web. Our research takes place in this context to make easy the obtaining of the organization definition from Web resources. In this paper, we present a definitional Question Answering (QA) system for the Arabic language called DefArabicQA. This system outperforms the use of Web searching by two criteria: (i) permits to ask by an ordinary question (e.g., ‘What is X?’) instead of asking by keywords query; (ii) returns an accurate answer instead of mining the Web searching results in order to find the expected information.

The paper is organized as follows: Section 2 provides an overview of the Arabic QA systems. Section 3 presents our definitional QA system DefArabicQA. Section 4 presents the realized experiments and Section 5 discusses the obtained results. A conclusion and some future directions for our work are exposed in Section 6.

2 Related works

QA systems are designed to retrieve the exact answers from a set of knowledge resources to the user question. Many researches are interested in this task in many competitions (e.g., TREC\(^1\), CLEF\(^2\) and NTCIR\(^3\)). An analysis of the TREC QA task experiments shows that two kinds of questions are mainly involved: factual and definition questions. A factual question is a simple fact retrieval where the answer is often a named entity (e.g. ‘Who is the president of the League of Arab States?’). Whereas a definition question is a question asking for any important information about someone or something (e.g., ‘What is the League of Arab States?’). Unfortunately, the evaluation platforms of QA task in the mainly evaluation conferences do not include the Arabic language. To our knowledge, no research has been done on Arabic definitional QA systems. However, there are some attempts to build factual QA systems (e.g. Hammo et al.,2002; Benajiba et al.,2007a; Brini et al.,2009). We cited below an overview of these factual Question Answering systems. (Hammo et al., 2002; 2004) developed QARAB a factual QA system. They employed information retrieval techniques to identify candidate passages, and sophisticated natural language processing techniques to parse the question and the top 10 ranked passages. They adopted a keyword matching strategy to identify answers. The answer identified is the whole sentence matching the question keywords. The evaluation process of this system was based on 113 questions and a set of documents collected from the newspaper Al-Raya. They obtained a precision equal to 97.3%, recall equal to 97.3% and MRR equal to 0.86 (Hammo et al.,2004). The average length of the answers obtained was 31 words. (Kanaan et al.,2004) developed a QA system using approximately the same method of (Hammo et

\(^1\) Text Retrieval Conference \texttt{http://trec.nist.gov/}

\(^2\) Cross-Language Evaluation Forum \texttt{http://clef-campaign.org/}

\(^3\) NII Test Collection for IR Systems \texttt{http://research.nii.ac.jp/ntcir/}
al., 2002) system’s. Their evaluation was based on a set of 25 documents from the Web and 12 questions. (Benajiba et al., 2007a) developed ‘ArabiQA’ a factual QA system. They employed Arabic-JIRS\(^4\) (Benajiba et al., 2007b), a passage retrieval system to search the relevant passages. They used also the named entity system ANERsys (Benajiba et al., 2007c) to identify and classify named entities within the passages retrieved. The test-set consists of 200 questions and 11,000 documents from Wikipedia Arabic version. They reached a precision of 83.3% (Benajiba et al., 2007a). (Brini et al., 2009) developed a prototype to build an Arabic factual Question Answering system using Nooj platform\(^5\) to identify answers from a set of education books. Most of these researches cited above, have not made test-bed publicly available, which makes it impossible to compare their evaluation results.

As we have already said, there is not a research focused on definitional QA systems for the Arabic language. Therefore, we have considered that an effort needs to be done in this direction. We built an Arabic QA system, which we named DefArabicQA that identifies and extracts the answers (i.e., exact definitions) from Web resources. Our approach is inspired from researches that have obtained good results in TREC experiments. Among these researches we cite the work of (Grunfeld & Kwok, 2006) which is based on techniques from IR, pattern matching and metakeyword detection with little linguistic analysis and no natural language understanding.

3 The DefArabicQA system

The architecture of the DefArabicQA system is illustrated in Figure 1. From a general viewpoint, the system is composed of the following components: i) question analysis, ii) passage retrieval, iii) definition extraction and iv) ranking candidate definitions.

This system does not use any sophisticated syntactic or semantic techniques, as those used for factual QA systems (Hammo et al., 2002; Benajiba et al., 2007).

3.1 Question analysis

This module is a vital component of DefArabicQA. The result of this module is the identification of the topic question (i.e., named entity) and the dedication of the answer type expected. The question topic is identified by using two lexical question patterns (Table. 1) and the answer type expected is deduced from the interrogative pronoun of the question.

<table>
<thead>
<tr>
<th>Question patterns</th>
<th>Expected answer types</th>
</tr>
</thead>
<tbody>
<tr>
<td>من هو &lt;topic&gt;?</td>
<td>Person</td>
</tr>
<tr>
<td>ما هو &lt;topic&gt;?</td>
<td>Organization</td>
</tr>
</tbody>
</table>

Table 1. Question patterns and their expected answer types used by DefArabicQA system

3.2 Passage retrieval

The passage retrieval module collects the top-n snippets retrieved by the Web search engine. This specific query is constituted of the question topic which is identified by the question analysis module. After collecting the top-n snippets, only those snippets containing the integrate question topic are kept on the basis of some heuristic (e.g. length of a snippet must be more than 13 characters).

3.3 Definition extraction

This module is the core module of a definitional QA system and it is composed of two sub-modules that are in charge of: i) identifying candidate definitions, and ii) filtering candidate definitions.

3.3.1. Identifying candidate definitions

In this step, we identify and extract candidate definitions from the collection of snippets collected in the passage retrieval module. We use lexical patterns to identify these candidate definitions. Generally, a lexical pattern is a sequence of strings (e.g., words, letters and punctuation symbols) which provide a context to identify the exact answers. It reflects a common use of written styles used to introduce an organization.

In our context, patterns are created manually and no natural language processing is employed in their construction. A candidate definition is identified by a specific pattern if the surrounding of the question

---

\(^4\) http://sourceforge.net/projects/jirs/
\(^5\) http://www.nooj4nlp.net/pages/nooj.html
topic in a snippet is recognized by a specific pattern.

3.3.2. Filtering candidate definitions
We use heuristic rules to filter the identified candidate definitions. These heuristic rules are deduced from the observation of a set of annotated candidate definitions (i.e., a collection of candidate definitions divided into incorrect candidate definitions and correct candidate definitions).

3.4 Definition ranking
The component “definition ranking” is based on a statistical approach. We used a global score to rank candidate definitions retained in the “Definition Extraction” module. This global score is a combination of three scores related to three criteria of a candidate definition: i) pattern weight criterion, ii) snippet position criterion, and iii) word frequency criterion. We present to the user the first top-5 candidate definitions ranked according to their global scores.

3.4.1. Pattern weight criterion \((C_1)\)
The score of this criterion is the weight of the pattern that has identified the candidate definition \(CD_i\). This score is represented by:

\[
C_1(CD_i) = w_i
\]

Where \(w_i\) presents the weight of pattern \(i\). We associate a weight to each pattern according to its relevance.

3.4.2. Snippets position criterion \((C_2)\)
The score of this criterion represents the position of the snippet that contains the candidate definition (in the snippets collection). This score is done by:

\[
C_2(CD_i) = p_i
\]

Where \(p_i\) is the snippet position containing the candidate definition \(CD_i\).

3.4.3. Word frequency criterion \((C_3)\)
The score of this criterion represents the sum of the frequencies of the words occurring in a candidate definition. According to this criterion, the candidate definition \(CD_i\) score is calculated as follows. Firstly, we construct a centroid vector containing common words across candidate definitions with their frequencies, beyond stopwords. Secondly, we calculate the frequency sum of the words recurring in both \(CD_i\) and centroid vector as indicated by the following formula:

\[
C_3(CD_i) = \sum_{k=1}^{n} f_{ik}
\]

Where \(n\) is the number of words which occur in the centroid vector and in the candidate definition \(CD_i\), \(1 \leq k \leq n\) and \(f_{ik}\) is the frequency of word \(k\).

3.4.4. Criteria aggregation
In order to aggregate the three criteria described above, we first proceed to the normalization of the score of each criterion by dividing it by the maximum score as follows:

\[
C_{i,j} = \frac{C_{i,j}}{\text{Max}C_i}
\]

Where \(i\) is a candidate definition and \(j\) a criterion. Then, we combine the three normalized scores in order to obtain the global score \(GS\) of the candidate definition \(CD_i\). This global score is obtained by:

\[
GS(CD_i) = \sum_{j=1}^{3} C_{i,j}
\]

4. Experiments and results
This section describes two experiments carried out using the DefArabicQA system. The first experiment was carried out using Google Search engine\(^6\), while the second experiment was carried out using Google Search engine and the free encyclopedia Wikipedia Arabic version\(^7\). In both experiments, we used 50 organization definition questions\(^8\) similar to these used in TREC. The system was assessed by an Arabic native speaker. As evaluation metrics, we use MRR. It is a measure used in TREC QA section and 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 (Voorhees, 2001).

4.1 Results of the first experiment
Out of the 50 questions in the test collection 41 questions (82%) were answered correctly by complete definitions in the top-five candidate definitions. 54% of the questions were answered by the first candidate definition returned, 14% by the second candidate definition, 6% by the third candidate definition, 6% by the fourth candidate

---

\(^6\) http://www.google.com/intl/ar/
\(^7\) http://ar.wikipedia.org/w/index.php?title=خاصص_بحث&search=
\(^8\) Resources available for research purpose at: http://sites.google.com/site/omartrigui/downloads
definition, 2% by the fifth candidate definition as shown in Table 2. The systems missed 18% of the questions as shown in Table 3. MRR was equal to 0.70 as shown in Table 4.

4.2 Results of the second experiment
The main goal of the second experiment is to measure the value added by the Web resource Wikipedia to the results obtained in the first experiment with the Google search engine. In this experiment, we used the same set of questions of the first experiment with Google search engine and Wikipedia as Web resources. Out of the 50 questions in the test collection, 45 questions (90%) were answered correctly by complete definitions in the top-five candidate definitions. 64% of the questions were answered by the first returned candidate definition, 16% by the second candidate definition, 4% by the third candidate definition, 2% by the fourth candidate definition and 4% by the fifth candidate definition as shown in Table 2. The system missed 10% of the questions as shown in Table 3. The obtained value of MRR is 0.81 (see Table 4).

<table>
<thead>
<tr>
<th>Rank</th>
<th>Experiment I</th>
<th>Experiment II</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>27 (54%)</td>
<td>32 (64%)</td>
</tr>
<tr>
<td>2nd</td>
<td>7 (14%)</td>
<td>8 (16%)</td>
</tr>
<tr>
<td>3rd</td>
<td>3 (6%)</td>
<td>2 (4%)</td>
</tr>
<tr>
<td>4th</td>
<td>3 (6%)</td>
<td>1 (2%)</td>
</tr>
<tr>
<td>5th</td>
<td>1 (2%)</td>
<td>2 (4%)</td>
</tr>
<tr>
<td>Top-five</td>
<td>41 (82%)</td>
<td>45 (90%)</td>
</tr>
</tbody>
</table>

Table 2. Rate of the answered questions for each Rank (the Top-5 positions)

<table>
<thead>
<tr>
<th></th>
<th>Experiment I</th>
<th>Experiment II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top-5</td>
<td>9 (18%)</td>
<td>5 (10%)</td>
</tr>
</tbody>
</table>

Table 3. Rate of non answered questions (in the Top-5 positions)

<table>
<thead>
<tr>
<th></th>
<th>Experiment I</th>
<th>Experiment II</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRR</td>
<td>0.70</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Table 4. MRR values for both experiments

5. Discussion
The two experiments cited above showed that our approach applied in DefArabicQA system returned reasonably good results. The Web resource Wikipedia has improved the results of DefArabicQA when it was coupled with Google in the second experiment. The MRR was increased from 0.70 (in the first experiment) to 0.81 (in the second experiment) and the rate of non answered question in the Top-5 positions was decreased from 18% (in the first experiment) to 10% (in the second experiment). Also, the Rate of the questions answered by the first returned candidate definition was increased from 54% (in the first experiment) to 64% (in the second experiment).

6. Conclusion and future work
In this paper we proposed a definitional Question Answering system called DefArabicQA. This system provides effective and exact answers to definition questions expressed in Arabic language from Web resources. It is based on an approach which employs a little linguistic analysis and no language understanding capability. DefArabicQA identifies candidate definitions by using a set of lexical patterns, filters these candidate definitions by using heuristic rules and ranks them by using a statistical approach.

Two evaluation experiments have been carried out on DefArabicQA. The first experiment was based on Google as a Web resource and has obtained an MRR equal to 0.70 and a rate of questions answered by the first answer equal to 54%, while the second experiment was based on Google coupled with Wikipedia as Web resources. In this experiment, we obtained an MRR equal to 0.81 and a rate of questions answered by the first answer equal to 64%. 50 definition questions are used for both experiments.

As future works, we plan to improve the quality of the definitions when it is truncated. Indeed, in some cases, few words are missed at the end of the definition answer. This is due to the fact that the snippet itself is truncated. As a solution, we will download the original Web page and segment the useful snippet correctly using a tokenizer. We also plan to conduct an empirical study to determine different weights to the three used criteria for ranking the candidate definitions. These weights will reflect the importance of each criterion.

Acknowledgments
This research work started thanks to the bilateral Spain-Tunisia research project on "Answer Extraction for Definition Questions in Arabic" (AECID-PCI B/017961/08).

The work of the third author was carried out in the framework of the AECID-PCI C/026728/09 and the TIN2009-13391-C04-03 research projects.

References


