Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 04, No. 02, August 2018.

TOWARD AN ARABIC QUESTION ANSWERING SYSTEM OVER LINKED DATA

Abdelghani Bouziane[1], Djelloul Bouchiha[1], Noureddine Doumi[2] and Mimoun Malki[3]

(Received: 31-Dec.-2017, Revised: 21-Mar.-2018 and 09-Apr.-2018, Accepted: 15-Apr.-2018)

ABSTRACT

The increasing interest in Arabic natural language processing and semantic Web research involves an emerging need to the development of new Question Answering Systems (QAS). These systems allow users to ask a question in Arabic natural language and get the relevant answer. However, most existing QA systems focused on English and Latin-based languages. Less effort has been concentrated on the Arabic language, which belongs to "Semitic Languages". This work is an early version of a new domain-independent Arabic question answering system over linked data, which aims to particularly help Arab users to explore the Arabic Semantic Web based on Arabic ontology. We describe with sufficient details the different modules of our proposed system, which uses language parser, finite state automaton and semantic Web techniques to linguistically process and answer Arabic natural language question. Experiments have been carried out to evaluate and show efficiency of the proposed system.

KEYWORDS

Question answering system, Natural language processing (NLP), Linked data, Arabic language, Semantic Web, Ontology, SPARQL.

1. INTRODUCTION

Arabic is the most spoken language in the Semitic group and the official or co-official language of 26 countries, spoken by more than 422 million people in the Middle East and North Africa. Arabic is the fourth top language in the internet with 185 million users [1]. 54% of Google searches in the Middle East and North Africa (MENA) are now made in Arabic, 34% in English and 8% in French [2].

This accumulation of Arabic information on the Web, available in large quantities and in various formats, includes structured and unstructured data. This large amount of data must be accessible and controlled by users who want to access and manipulate this information. So, more sophisticated systems are needed. For this need, the existing search tools, like search engines and query languages, make finding information a complex and expensive task in terms of time. This difficulty has motivated the development of new adapted search tools; namely, Question Answering Systems (QAS), which allow users to ask a question in natural language and get the relevant answer.

Currently, mutation of these systems to the Web of data seems necessary to find the correct and accurate answers to users' questions. New question answering systems have to deal with Linked Data instead of the Web of documents. The Linked Data initiative aims at publishing structured and interlinked data on the Web by using Semantic Web technologies [3]. These technologies provide different languages for expressing data as graphs (RDF) and querying it (SPARQL) [4].

Despite the considerable research in question answering systems over semantic Web for English language, the development of Arabic QA systems over semantic Web is at its nascent stage. Due to the limits of Arabic NLP technologies [5] in Arabic Semantic Web, there are few works in the field of Arabic QAS over semantic Web and ontology [6]. Motivated by this challenge, we present in this paper an Arabic question answering system over linked data. The main objective of our system is to provide exact answers for Arabic natural language factoid questions, asked by native users, who don't know the complicated SPARQL query language. First, the proposed system receives as input an

1. A. Bouziane and D. Bouchiha are with EEDIS Lab., Ctr. Univ. Naama, Inst. Sciences and Technologies, Dept. Mathematics and Computer Science, Algeria. Emails: abdelghani.bouziane@univ-sba.dz and djelloul.bouchiha@univ-sba.dz
2. N. Doumi is with University of Saida, Dept. of Computer Science, Algeria. Email: noureddine.doumi@univ-saida.dz
3. M. Malki is with Lab RI- Sidi Bel-Abbes Ecole Supérieure en Informatique de Sidi Bel-abbes, Algeria. Email: m.malki@esi-sba.dz

Arabic natural language question, from which it provides the resource and keywords to the first module. Then, it uses a general SPARQL query to explore the ontology and gives the exact predicate by matching terms between keywords and predicates. Finally, our system provides an answer to the user question by running a final appropriate SPARQL query.

The rest of the paper is organized as follows. Section 2 highlights the challenge of the development of Arabic question answering systems over semantic Web. Section 3 reviews the existing related work. Section 4 presents the architecture of our proposed system. An illustrative example is covered in Section 5. Section 6 evaluates the resource extraction method, the stop-words' removal algorithm and the overall system accuracy. Finally, we summarize the paper and highlight the future work directions in Section 7.

2. MOTIVATION AND CHALLENGE

In this study, we present an Arabic natural language question answering system over semantic Web. The issue of developing such a system for poor resources in Arabic NLP [7] and Arabic Semantic Web is very challenging. Let's discuss in the following part of this section the difficulties in Arabic NLP and Arabic Semantic Web to highlight the development of Arabic question answering systems over semantic Web.

Arabic language is characterized by a relatively complex morphology. It has a rich system of morphological inflection. Arabic has also a high degree of ambiguity resulting from its diacritic-optional writing system, common deviation from spelling standards and absence of capital letters [8]. The existing tools and resources deal mainly with the Modern Standard Arabic (MSA), but the Arab World uses different Arabic dialects such as: (Egyptian Arabic (EGY), Levantine (LEV), Gulf Arabic (GLF), North African (Maghrebi) Arabic (Mag), Iraqi Arabic (IRQ), Yemenite Arabic (Yem)) [9]. There are limited resources and tools for Arabic language processing compared to English language. The other difficulty is the lack of Arabic support for Arabic language in the semantic Web technologies. The Arabic Semantic Web is very far from the best performance compared to English semantic Web. The difficulties of Arabic natural language processing hinder the development of Arabic Semantic Web, because the NLP is an important component of the semantic Web [10].

Another important component of the semantic Web is the Arabic ontology [11], which is said to be the foundation of the creation of Arabic Semantic Web designs. Recently, several works have dealt with the semantic Web for the Arabic language [11]-[12]. As can be seen from Table 1, there is weakness in ontology management tools for the Arabic language. So, there is a need to develop controlled vocabulary and ontology management tools for Arabic ontology for the foundation of Arabic Semantic Web [13].

Table 1. Examples of semantic web tools and their support to the Arabic language [13].

Tool	Arabic RDF ontology	Arabic OWL ontology	Arabic Query	Description
Protege	Support	Limited support	Limited support	Ontology editor
Jena	Support	Support	Limited support	Reasoner and processors
Sesame	Limited support	Limited support	No support	RDF database
KAON2	No support	No support	No support	Reasoner and ontology management

Table 1 shows how much some semantic Web tools support the Arabic language. The first column of Table 1 corresponds to the tool name. The columns "Arabic RDF ontology" and "Arabic OWL Ontology" indicate whether the tool supports these two languages (RDF and OWL) to express an Arabic ontology. The column "Arabic Query" indicates whether the tool allows interrogating the Arabic ontology with Arabic question. The "Description" column indicates the tool function.

3. RELATED WORK

The first survey of question answering systems backs to 1965, in which several systems were reviewed for the English language during the previous five years [14]. One of the first subtasks of question

answering systems is the Natural Language Interfacing to Data Bases (NLIDB), such as BASEBALL [15], PARRY [16], LUNAR [17] and NLWIDB [18]. Research in QA received a big surge in interest when a shared task on factoid QA was included in the 8th Text REtrieval Conference (TREC) [19]. Most systems process textual information, such as Youzheng et al. [20], Mulder [21], PALANTIR [22], QALC [23], Gómez et al. [24], Ryu et al. [25]. These question answering systems can be divided into three main distinct subtasks [26]-[28], which are Question Analysis, Document Retrieval and Answer Extraction. Most question answering systems follow these three subtasks. However, in the last years, the mutation of these systems to the Web of data seems necessary to find the correct and accurate answers to questions. New question answering systems have to deal with Linked Data instead of the Web of documents. Several systems appeared, such as AquaLog [28], SWIP [29], Xser [30], DeepQA IBM Watson's system [31] and E-librarian [32].

In the field of question answering systems for Arabic language, the situation is less bright. Research in this field is slow and gives limited results for all subtasks of QAS due to the lack of resources and tools in Arabic NLP [6]. Next are given works on Arabic QAS.

In [33], the authors proposed a question analysis for Arabic question answering systems using Stanford POS Tagger & parser for Arabic language, named entity recognizer, tokenizer, Stop-word removal, Question expansion, Question classification and Question focus extraction components to retrieve and extract the correct answer.

Al-Bayan [34] is a Question answering system for the holy Quran. The user asks Arabic natural language questions about the Quran. The system first retrieves the most relevant Quran verses. Second, it extracts the passage that contains the answer from two sources; the Quran and its interpretation books (Tafseer).

AQuASys [35] is designed to answer Arabic fact-based questions. It is composed of three modules: a question analysis module, a sentence filtering module and an answer extraction module.

In [36], the information source of the system is a given corpus or the Web pages. It uses supervised support vector machine (SVM) classifier for question classification and answer selection, to generate the exact answer for a given question in Arabic natural language.

AR2SPARQL [37] is an Arabic natural language interface for semantic Web, which uses linguistic and semantic analysis to convert the Arabic query into RDF triples, which are then matched with ontology triples to retrieve an answer.

AQAS [38] extracts answers from structured data. It is the first system of its type for the Arabic language. Knowledge from the radiation domain is presented using the frame technique. There is no published evaluation about AQAS.

QARAB [39]; [51] is an unconnected (non-Web-based) QA system for only factoid questions. No other type of question is supported. It uses IR and NLP techniques to extract answers from a collection of Arabic newspaper texts.

In [40], the authors use tagging rules and question patterns to analyze and understand an Arabic question in Arabic question answering environment.

ArabicQA [41] is an Arabic question answering system based on a Passage Retrieval (PR) module, a Named Entities Recognition (NER) module and an answer extraction module for Arabic texts.

In [42], the author describes a way to access Arabic Web Question Answering (QA) corpus using a chatbot (ALICE open-source chatbot initiative). ALICE is the Artificial Linguistic Internet Computer Entity. The system uses a simple (but large) set of pattern-template matching rules and converts a text corpus into the AIML chatbot model format.

In [43], the authors use lexical pattern for defining questions to extract the focus and Wikipedia article infoboxes, to generate cooperative answers for user-definitional questions. This approach can be integrated in all question answering systems.

The following table summarizes related works according to four criteria: (1) "Target source" indicates the name and type of the analyzed data, (2) "Question analysis techniques" correspond to the used analysis techniques by the considered system, (3) "Question classification techniques" correspond to

the used classification techniques and (4) "Performance", which gives some experimental results.

Table 2. Some Arabic QAS features and techniques.

System	Target source	Question analysis techniques	Question classification techniques	System Performance
Al bayan [34]	Quran and interpretation books (unstructured)	NLP technique	SVM classifier	Experts evaluation results: 0.73%
AQuASys [35]	Documents (unstructured)	Defined question structures and NLP technique	Defined question types and forms	Precision: 66.25 % Recall: 97.5%
AQAS [38]	Knowledge bases radiation diseases domain (structured)	Parser: morphological analysis	No	Not mentioned
QARAB [51]	Al-Raya Arabic newspaper text (unstructured)	NLP technique	Using a set of known question type	Precision: 97.3% Recall: 97.3%
WAHEED [36]	Web of document (unstructured)	NLP technique	SVM classifier	Mean Reciprocal Rank (MRR) : 65%
AR2SPARQL [37]	Ontology (structured)	NLP technique	No	Precision: 85.24% Recall: 61.61% F-measure: 71.5%
Al-Shawakfa [40]	Corpora of documents (unstructured)	NLP technique	Defined question type	Accuracy: 78.15% Recall: 97% F-measure: 86.56%
ArabicQA [41]	Set of documents (unstructured)	NLP technique	Defined question type	Not mentioned
Abu shawar [42]	Text corpus (unstructured)	Pattern-template matching rules	No	Recall: 93%
DefArabicQA [43]	Wikipedia article (semi-structured)	Lexical pattern	No	Accuracy: 63%

Most of works in Arabic question answering systems deal with unstructured data. Currently, with the growth of linked data technologies, new systems must interact with linked data instead of the Web of documents. Our proposed system tries to address these challenges by proposing a new solution that helps answer questions in Arabic natural language by translating them into SPARQL queries. In the next section, we describe the architecture of our proposed system.

4. PROPOSED SYSTEM

We propose an Arabic question answering system that transforms Arabic factoid questions into SPARQL query and then provides an exact answer drawn from an Arabic ontology-based knowledge base. Our proposed system accepts the Arabic simple factoid questions which may be in the following format (من هو\man. huw~a*who is*, من هي\man. hiy~a*who is*, ما هو\maA huw~a*what is*, ما هي\maA hiy~a*what is*, متى\mataý\when, مما\mim~aA*from what*, أين\Âay.na*where*, كيف\kay.fa*how*, في أي\fiy Âay~u*in what*, كم\kam.*how many*, بأي\biÂay~i*in what* …).

First, through its interface, the system receives the question in NL, processes it and finally produces an answer after formulating a SPARQL query that can be executed on Arabic Semantic Web, based on Arabic ontology. The transformation process is composed of three consecutive modules. Each one is composed of multiple steps depicted as presented in Figure 1. It starts with question processing. Then, predicate recognition is done. Finally, SPARQL query is formulated and executed.

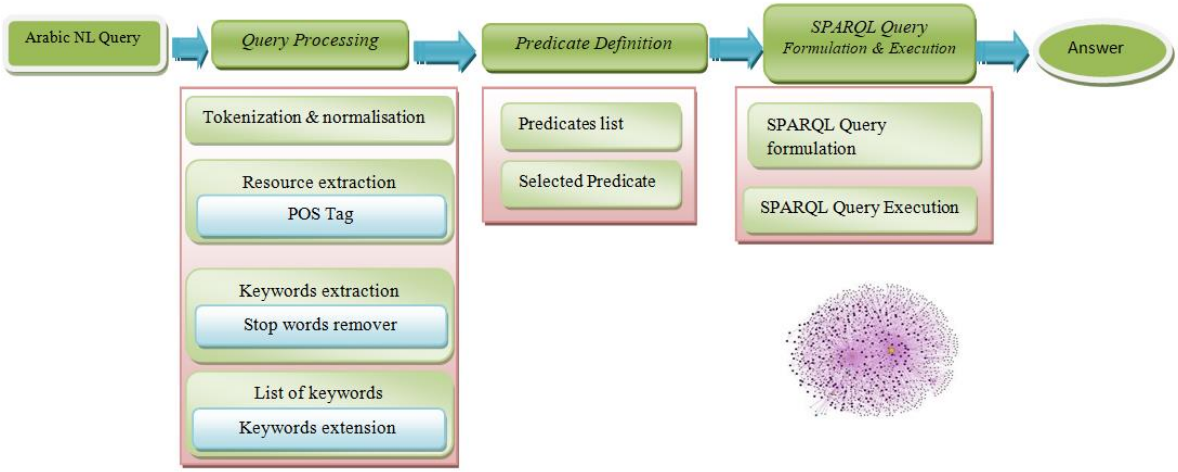

Figure 1. Architecture of the proposed system.

In the next sub-sections, we explain the above mentioned modules.

4.1 Question Processing Module

It is an important and crucial module. It allows analyzing an input natural language question, in order to have its constituents. This module has a big impact on the accuracy and the performance of any QA system. In our proposed system, the input question is linguistically processed and analyzed by the Question processing module, which consists of four steps: Tokenization and Normalization, Resource Extraction, Keywords Extraction and getting Keywords List. As a result, it provides the keywords list and the resources (named entities).

4.1.1 Tokenization and Normalization

The common step in NLP is Tokenization, which denotes the segmentation of the natural language text (question in our case) into individual consecutive basic units. A step of word normalization is necessary to reduce the spelling errors. These errors appear because the Arabic letter can be written in different styles. The correction of the most common spelling errors involves normalization of Arabic Alif 'ا' and Ya 'ي' characters [9]. We use MADAMIRA tools for the Tokenization and Normalization step [8].

Example: The Tokenization and Normalization step of the Arabic natural language question: ما هي عاصمه ألجزائر؟\maA hiy çaASimah ÂljazaAŷir ?*what is the capital of Algeria* ? provides "ما، هي، عاصمة، الجزائر". It normalizes the Alf 'ا'and Taa 'ة' in the words of this NL question.

Note that in this example and in the rest of this paper, every time we give an Arabic text and for readability purposes, we follow it with its HSB transliteration [44] and English translation for readability purposes.

4.1.2 Resource Extraction

The named entity is very important in most question answering systems for structured or unstructured data [45]. To date, as per to our knowledge, there are no Arabic Named Entity Recognition (NER) systems available for free. So, to gap this difficulty, we implemented our own NER. In our system, the named entity in the input question is the target resource to be explored for getting the correct answer. So, the Resource Extraction process consists of extracting the last Nominal Phrase (NP) of the Arabic question from the parse tree of the considered question. We use the Stanford Part-Of-Speech Tagger (pos-tagg) which is a Java implementation designed to provide a simple description of the grammatical relationships in a sentence [46].

Example: The last NP of the question ما هي عاصمة الجزائر؟\maA hiy çaASimaħu AljazaAŷir ?\ is the named entity الجزائر\AljazaAŷir*Algeria* in Figure 2.

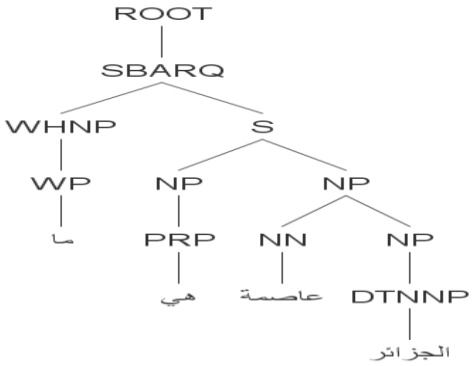

Figure 2. Parse tree of an Arabic question, given by Stanford POS Tagger.

4.1.3 Keywords Extraction

The keywords of the input question are used to generate the predicate of the <Subject, Predicate, Object> triple and to formulate the final SPARQL query.

To extract keywords from the input question, the system processes in two steps: First step is stop-words removal. Stop-words are the noisy words which frequently occur in the Arabic NL questions, such as prepositions, conjunctions and interrogative words. We propose a Finite State Automaton (FSA) that recognizes stop-words to be removed in our input Arabic question. The finite state automata technique (Figure 3) can accelerate the stop-words removing process [47].

In order to extract keywords, the second step consists of removing the words already recognized in the Resource Extraction step. Keywords are words that are not removed.

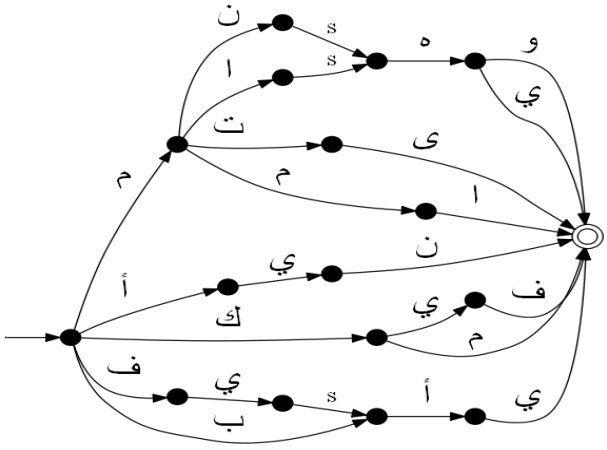

Figure 3. Stop-words finite sate automaton.

Example: From the text ما هي عاصمة الجزائر\maA hiy çaASimaħu AljazaAŷir*what is the capital of Algeria*, the stop-words removal gives ما هي \maA hiy*what is the*. Then the resource is الجزائر\AljazaAŷir*Algeria*. The rest عاصمة\çaASimaħu*capital* is the keyword of the question in natural language.

4.1.4 Keywords Extension

Keywords extracted in the previous step and the predicate used in the ontology may have a different morphology, but the same meaning. To gap this difficulty, we use Arabic WordNet [48, 49] to find synonyms, in order to extend the keywords set. These synonyms are morphologically normalized by the normalization process used in the first step of this module. A list of keywords is built to increase the probability to define the exact predicate in the next module.

4.2 Predicate Recognition

After the resource was fixed by the question processing module, the system uses a simple SPARQL

query to get all properties of our resource.

Question 1: ما هي عاصمة الجزائر؟\maA hiy ςaASimaħu AljazaAŷir*what is the capital of Algeria ?*

Question 2: متى ولد هواري بومدين؟ \matý wulida huwaAriy buwmad.yan*what is the birthday of Houari Boumedienne ?*

Resources of questions 1 and 2 are الجزائر\AljazaAŷir*Algeria* and هواري بومدين\huwaAriy buwmad.yan*Houari Boumedienne*, respectively. Now, we have to build a SPARQL query to retrieve the DBpedia Arabic properties of the resource in question:

Select ?p where{ < arabic-ontology-2#الجزائر > ?p ?o}
Select ?p where{ < arabic-ontology-2#هواري بومدين > ?p ?o}

The SPARQL resulting file is an XML document that is valid with respect to the XML Schema. The principal element is the SPARQL one. Inside the SPARQL element, there are two sub-elements: head and a result element. The result element contains the complete sequence of the query result. It has a child binding element. In this binding element, we find the query variable's name and value. We use the XQuery language to explore the XML document and build the predicate list.

Now, we match the keywords list with the predicates list. It is a lexical matching process, where the common word is selected to be the predicate.

4.3 SPARQL Query Formulation and Execution

Now, we are able to formulate the final SPARQL query with the resource and predicate to provide the exact answer from the Arabic ontology. For this, we use the template query:

Select ? Object where {Resource Property ?Object}

For the question ما هي عاصمة الجزائر؟\maA hiy ςaASimaħu AljazaAŷir ?*what is the capital of Algeria ?*, the SPARQL query is:

Select ? Object where {< arabic-ontology-2#الجزائر> *onto:* عاصمة *?Object}*

The result of the SPARQL query is an XML document, which will be parsed by the same process of the previous step to extract the correct answer.

5. ILLUSTRATIVE EXAMPLE

Now, we show a complete illustrative example of processing an Arabic natural language question. In this example, we assume that the introduced question to our system is:

من هو مؤلف رياض الصالحين؟\man huwa muŵlif riyaAD AlSAliHiyn?\, which corresponds to the English question *who is the author of The Meadows of the Righteous ?*

The question processing module takes place with 4 steps: Tokenization and Normalization, Resource Extraction, Keywords Extraction and Keywords List. Finally, it produces the following

Question: من هو مؤلف رياض الصالحين؟
Tokenization and Normalization: من, هو, مؤلف, رياض, الصالحين
Resource Extraction: the produced parse tree is (ROOT (SBARQ (WHNP (WP من)) (S (NP (PRP هو)) (NP (NN مؤلف) (NP (NNP رياض) (DTNNP الصالحين)))) (PUNC ?))). The resource is: رياض الصالحين
Keywords Extraction: first, the system eliminates the stop-words and the resource, so the keyword is: مؤلف
Keywords list: the system extends the keywords set by using Arabic WordNet, مؤلف\muŵalif\, كاتب\kaAtib\ *writer, author*.

The question processing module produces the following information to be the input of the next module.

Resource: رياض الصالحين\riyaAD AlSaAliHiyn*The Meadows of the Righteous*

List of keywords: مؤلف\muŵalif\, كاتب\kaAtib\ *writer, author*.

Predicate Recognition module: this module generates and executes the SPARQL query through the ontology to retrieve all predicates:

Select ?p where{ < arabic-ontology-2#رياض الصالحين > ?p ?o}

The result of the SPARQL query after processing is the predicate lists: اسم، مؤلف، لغة كتاب، بلد صدور\Ais.m, muŵalif, luγaħ kitaAb, baladu Suduwr*name, author, language of book, country of publication*.

Then, we select the predicate: مؤلف\ muŵalif*author* as a common term between the predicates list and the keywords list.

Finally, the system provides an XML document result by formulating and executing the SPARQL query:

SELECT ?object
WHERE {< semanticweb.org/ghani/ontologies/2017/6/untitled-ontology-2#رياض_الصالحين >
onto:مؤلف ?object }

The answer is extracted from the XML document:
يحيى_بن_شرف_النووي\yaH.yaý_b.nu_šaraf_Alnawawiy\

6. EVALUATION

Before evaluating the performance of our proposed system, it is important to briefly introduce the used dataset which is an Arabic ontology that we developed for this evaluation. The ontology depicted in Figure 4 covers notions surrounding the concept *Person*. An excerpt of the ontology shows the ontology classes (e.g. *Person*, *Occupation*, *Country* and *Place*), the object properties and the data type properties.

We carried out several experiments for a full evaluation of our system. We evaluated the performance of the Resource Extraction method, the Stop-words removal algorithm and the overall system accuracy.

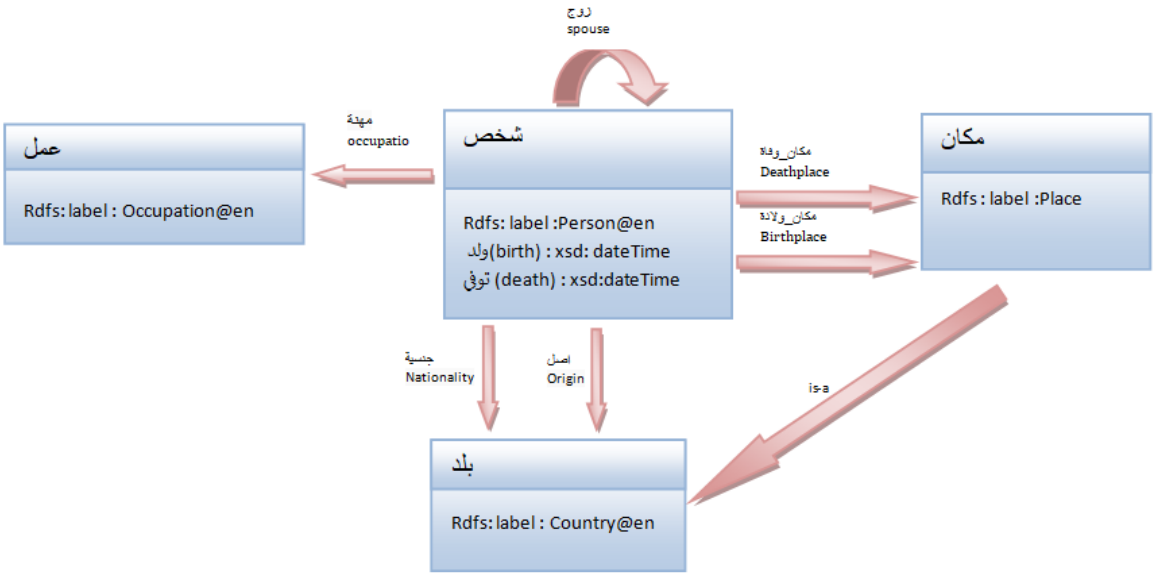

Figure 4. Excerpt of the ontology person.

For evaluating the resource extraction, the stop words removal and the overall system, we use a set of 50 factoid simple questions of different types about the ontology *Person*. This dataset is collected from the test-bed of the passage retrieval and question answering tasks proposed by Yassine Benajiba [50].

6.1 Evaluation Metrics

To measure the performance of our Resource Extraction, Stop-words Removal algorithm and the over-

all system, we used the following metrics: Precision, Recall and F-measure.

Precision (P) assesses the accuracy and is defined as follows:

Resource extraction process:

$$\text{Precision} = \frac{\text{number of correctely identified ressources}}{\text{total number of resources generated by the system}}$$

Stop words removal algorithm:

$$\text{Precision} = \frac{\text{number of correctely identified stop words}}{\text{total number of stop words generated by the system}}$$

System:

$$\text{Precision} = \frac{\text{number of correctely answered questions}}{\text{total number of questions answered by the system}}$$

The Recall (R) assesses the coverage and is defined as follows:

Resource extraction process:

$$\text{Recall} = \frac{\text{number of correctely identified ressources}}{\text{total number of resources}}$$

Stop-words removal algorithm:

$$\text{Recall} = \frac{\text{number of correctely identified stop words}}{\text{total number of stop words}}$$

System:

$$\text{Recall} = \frac{\text{number of correctely answered questions}}{\text{total number of introduced questions}}$$

Finally, F-measure is the trade-off, calculated by multiplying 2 times the product of Precision and Recall, divided by the sum of Precision and Recall. Mathematically, the F-measure formula is:

F-measure = 2 * (P*R) / (P+R).

6.2 Results and Discussion

In fact, the evaluations for the Resource Extraction, Stop-Words Removal algorithm and the overall system have been performed using dataset of [50]. The evaluation results are depicted in Table 3.

Table 3. Evaluation results, expressed in precision, recall and F-measure.

	Resource Extraction			Stop-Words Removal			System		
Number of questions	**Precision**	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
05	1	1	1	1	1	1	1	1	1
10	0,8	0,88	0,83	0,9	0,95	0,92	0,8	0,8	0,8
15	0,8	0,85	0,82	0,93	0,95	0,94	0,78	0,73	0,75
20	0,75	0,83	0,78	0,9	0,95	0,92	0,77	0,7	0,73
30	0,73	0,7	0,71	0,95	0,95	0,95	0,71	0,66	0,68
40	0,73	0,71	0,71	0,95	0,95	0,95	0,7	0,67	0,68
50	0,73	0,71	0,71	0,96	0,96	0,95	0,71	0,69	0,7

6.2.1 Resource Extraction

For the Resource Extraction process, the resulting Precision, Recall and F-measure start respectively from 0.73, 0.70 and 0.71 to 1, as shown in Table 3. Precision and Recall need improvement, which reflects the need to add grammar rules and gazetteers into the Recognition process. The failure cases can be explained mainly by the incorrect parsing of some Arabic questions. The Arabic resources, which are Arabic names, are often in the form of nominal phrases, which consist of a noun stem or an

adjective with nominal reference. In some cases, Arabic nouns are more complicated and consist of more than one nominal phrases or verbal phrases.

6.2.2 Stop-words Removal

Using our question dataset, the accuracy of the Stop-words Removal algorithm is illustrated in Table 3. The resulting Precision, Recall and F-measure for the entire 30 simple factoid questions are high (0.95), which projects high accuracy.

6.2.3 System

As can be seen in Table 3, the system achieves successfully an average of 0.66 in terms of Recall and 0.71 in terms of Precision. The most likely explanations of the failure results of our system are the following:

- Resource Extraction failure: The previous evaluation results showed that the Resource Extraction process has a limited result, which affects negatively the accuracy of the system as can be seen in Figure 5.

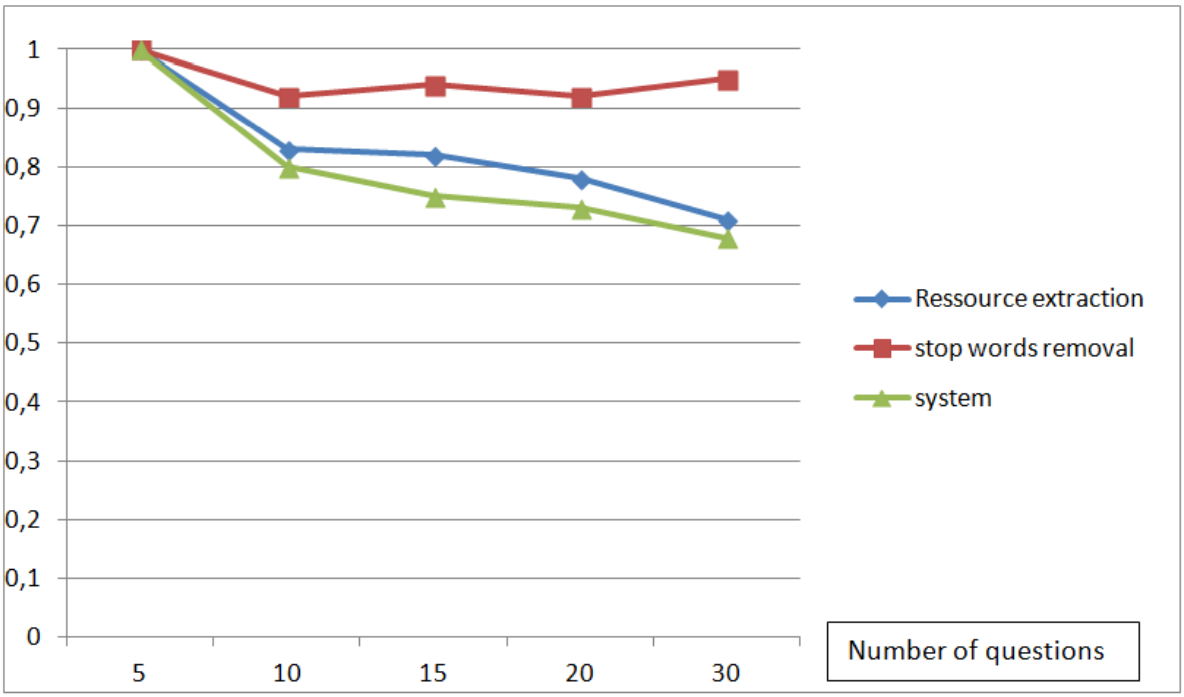

Figure 5. F-measure evaluation curves.

- Keywords failure: occurred when words could not be matched to any predicate of the appropriate resource. The most frequent causes of this type of errors can be classified into three main categories:
  - Recognition failure: this failure can occur when the user asks an Arabic question in unordinary and not standardized manner. Don't forget that our system deals with Modern Standardized Arabic (MSA) questions. The main cause of the Recognition failure can be explained by the Resource Extraction failure and the Stop-Words Removal failure.
  - Extension failure: occurs when the system doesn't generate any synonym of the predicate using the Arabic WordNet resource.
  - Matching failure: occurs when the list of keywords can't be matched to any predicate. So, the vocabulary of the ontology must use the most frequent words.

7. CONCLUSION AND FUTURE WORKS

We presented a question answering system that provides answers to questions expressed in Arabic natural language. We believe that the proposed system makes a step towards enabling users to explore the growing Arabic content on the Semantic Web. The system uses techniques from NLP and semantic Web to process the input question and to transform it into SPARQL query to get answers from Arabic Semantic Web based on ontology. First, the system applies a linguistic process to the input question in order to provide the resource and a list of keywords. Second, all the predicates of the resource are selected and matched to the keyword list in order to identify the appropriate predicate which form the <Resource, Predicate, Object> triple. Finally, our system formulates and executes a final SPARQL query to get an exact answer.

We discussed the major challenges of developing this kind of system for Arabic language, which is valuable for further study in more depth. Based on the evaluation results, it can be concluded that this research field is very promising. However, we plan to resolve some of the limitations we found in our evaluation by improving the existing modules using natural language processing techniques, as well as other tools and resources.

Since this is one of the first works aiming Arabic question answering system over linked data, there are many directions to extend our work. First is to add new modules such as question categorization. Second is to extend this work to cover other complex cases in Arabic question. Finally, is to apply our proposed approach to a real-world example of the Arabic DBpedia chapter, which is the real world example of the Arabic Semantic Web provided by the DBpedia community.

REFERENCES

[1] M. M. Goup, Internet World Users By Language: Top 10 Languages, [Online], Available: http://www.internetworldstats.com/stats7.htm, [Accessed October 2017].

[2] B. v. d. Beld, State of Digital, The Arabic Web: Numbers and Facts, General Statistics, [Online], Available: http://www.stateofdigital.com/the-arabic-web/, [Accessed October 2017].

[3] S. Albagli, R. Ben-Eliyahu-Zohary and S. E. Shimony, "Markov Network-based Ontology Matching," Journal of Computer and System Sciences, vol. 78, pp. 105-118, 2012.

[4] P. Hitzler, M. Krotzsch and S. Rudolph, Foundations of Semantic Web Technologies: CRC Press, 2009.

[5] W. Zaghouani, "Critical Survey of the Freely Available Arabic Corpora," arXiv Preprint arXiv:1702.07835, 2017.

[6] A. Bouziane, D. Bouchiha, N. Doumi and M. Malki, "Question Answering Systems: The Story Till the Arabic Linked Data," International Journal of Artificial Intelligence and Soft Computing (IJAISC), vol. 6, pp. 24-42, 2017.

[7] A. M. Ezzeldin and M. Shaheen, "A Survey of Arabic Question Answering: Challenges, Tasks, Approaches, Tools and Future Trends," Proceedings of The 13th International Arab Conference on Information Technology (ACIT 2012), pp. 1-8, 2012.

[8] A. Pasha, M. Al-Badrashiny, M. Diab, A. E. Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow and R. M. Roth, "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic," Ed. LREC2014, 2014.

[9] N. Habash, Introduction to Arabic Natural Language Processing: Morgan & Claypool, 2010.

[10] R. Guo and F. Ren, "Towards the Relationship Between Semantic Web and NLP," International Conference on Natural Language Processing and Knowledge Engineering, Dalian, 2009.

[11] M. M. Boudabous, L. H. Belguith and F. Sadat, "Exploiting the Arabic Wikipedia for Semi-automatic Construction of a Lexical Ontology," International Journal of Metadata, Semantics and Ontologies, vol. 8, pp. 245-253, 2013.

[12] H. Al-Feel, "The Roadmap for the Arabic Chapter of DBpedia," Mathematical and Computational Methods in Electrical Engineering, Proceedings of the 14th International Conference on Telecom. and Informatics (TELE-INFO '15), Sliema, Malta, pp. 115-125, 2015.

[13] M. Beseiso, A. R. Ahmad and R. Ismail, "A Survey of Arabic Language Support in Semantic Web," International Journal of Computer Applications, vol. 9, pp. 35-40, 2010.

[14] R. F. Simmons, "Answering English Questions by Computer: A Survey," Communications of the ACM, vol. 8, pp. 53-70, 1965.

[15] B. F. Green, A. K. Wolf, C. Chomsky and K. Laughery, "BASEBALL: An Automatic Question Answering," Proceedings of Western Joint Computer Conference, pp. 207-216, 1961.

[16] K. M. Colby, "Artificial Paranoia," Artificial Intelligence, vol. Vol. 2, 1971.

[17] W. Woods, R. Kaplan and B. Webber, "The Lunar Sciences Natural Language Information System," Cambridge, Massachusetts Final Report, 1972.

[18] A. Rukshan, R. Prashanthi and M. Sinnathamby, "Natural Language Web Interface for Database (NLWIDB)," Proceedings of the 3rd International Symposium, SEUSL, Oluvil, Sri Lanka, 2013.

[19] E. M. Voorhees and D. M. Tice, "The TREC-8 Question Answering Track Evaluation," NIST Special Publication 500-246: The 8th Text REtrieval Conference (TREC 8), 1999.

[20] W. Youzheng, H. Chiori, K. Hideki and K. Hisashi, "Leveraging Social Q&A Collections for Improving Complex Question Answering," Elsevier, Computer Speech and Language, vol. 29, pp. 1-19, 2015.

[21] C. Kwok, O. Etzioni and D. Weld, "Scaling Question Answering to the Web," Proceedings of the 10th International Conference on World Wide Web, Hong Kong, China, pp. 150-161, 2001.

[22] S. Harabagiu Dan, A. Harabagiu, D. Moldovan, C. Clark, M. Bowden, A. Hickl and P. Wang, "Employing Two Question Answering Systems in TREC 2005," Proceedings of the 14th Text REtrieval Conference, 2005.

[23] O. Ferret, B. Grau, G. Illouz, C. Jacquemin and N. Masson, "QALC - The Question-Answering Program of the Language and Cognition Group at LIMSI-CNRS," TREC-8, Columbia, 1999.

[24] J. M. G. Soriano, M. M. Y. Gómez, E. S. Arnal and P. Rosso, "A Passage Retrieval System for Multilingual Question Answering," International Conference on Text, Speech and Dialogue, pp. 443-450, 2005.

[25] P.-M. Ryu, M.-G. Jang and H.-K. Kim, "Open Domain Question Answering Using Wikipedia-based Knowledge Model," Information Processing & Management, vol. 50, pp. 683-692, 2014.

[26] R. Sutcliffe, A. Peñas, E. Hovy, P. Forner, Á. Rodrigo, C. Forascu, Y. Benajiba and P. Osenova, "Overview of QA4MRE Main Task at CLEF 2013," Working Notes CLEF, 2013.

[27] S. K. Ray and K. Shaalan, "A Review and Future Perspectives of Arabic Question Answering Systems," IEEE Transactions on Knowledge and Data Engineering, vol. 28, pp. 3169-3190, 2016.

[28] V. Lopez, V. Uren, M. Sabou and E. Motta, "Is Question Answering Fit for the Semantic Web? A Survey, " Semantic Web, vol. 2, pp. 125–155, 2011.

[29] C. Pradel, O. Haemmerlé and N. Hernandez, "Swip: A Natural Language to SPARQL Interface Implemented with SPARQL," 21st International Conference on Conceptual Structures (ICCS 2014), Iaşi, Romania, 2014.

[30] K. Xu, S. Zhang, Y. Feng and D. Zhao, "Answering Natural Language Questions *via* Phrasal Semantic Parsing," The Natural Language Processing and Chinese Computing, Third CCF Conference (NLPCC 2014), Shenzhen, China, 2014.

[31] A. Kalyanpur, B. K. Boguraev, S. Patwardhan, J. W. Murdock, A. Lally, C. Welty, J. M. Prager, B. Coppola, A. Fokoue-Nkoutche, L. Zhang, Y. Pan and Z. M. Qiu, "Structured Data and Inference in DeepQA," IBM Journal of Research and Development, vol. 56, pp. 10:1 - 10:14, May-June 2012.

[32] S. Linckels and C. Meinel, "A Simple Solution for an Intelligent Librarian System," Proceedings of the IADIS International Conference of Applied Computing (IADIS AC2005), Lisbon, Portugal, pp. 495-503, 2005.

[33] W. Ahmed and A. P. Babu, "Question Analysis for Arabic Question Answering Systems," International Journal on Natural Language Computing (IJNLC), vol. 5, December 2016.

[34] H. Abdelnasser, R. Mohamed, M. Ragab, A. Mohamed, B. Farouk, N. El-Makky and M. Torki, "Al-Bayan: An Arabic Question Answering System for the Holy Quran," Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP), Doha, Qatar, pp. 57–64, 2014.

[35] S. Bekhti and M. Al-Harbi, "AQuASys: A Question-Answering System for Arabic," Proceedings of the 13th International Conference on Applied Computer Science (ACS '13), Proceedings of the 2nd

International Conference on Digital Services, Internet and Applications (DSIA'13), Morioka City, Iwate, Japan, pp. 130-139, 2013.

[36] W. Ahmed, A. Pv and A. P. Babu, "Web-based Arabic Question Answering System using Machine Learning Approach," International Journal of Advanced Research in Computer Science, vol. 8, pp. 40-45, Jan./Feb. 2017.

[37] I. Al-Agha and A. Abu-Taha, "AR2SPARQL: An Arabic Natural Language Interface for the Semantic Web," International Journal of Computer Applications, vol. 125, 2015.

[38] F. A. Mohammed, K. Nasser and H. M. Harb, "A Knowledge-based Arabic Question Answering System (AQAS)," ACM SIGART Bulletin, vol. 4, pp. 21-30, Oct. 1993.

[39] B. Hammo, H. Abu-Salem and S. Lytinen, "QARAB: A Question Answering System to Support the Arabic Language," Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages (SEMITIC'02), Philadelphia, Pennsylvania, pp. 1-11, 2002.

[40] E. Al-Shawakfa, "A Rule-based Approach to Understand Questions in Arabic Question Answering," Jordanian Journal of Computers and Information Technology, vol. 2, pp. 210-231, 2016.

[41] Y. Benajiba, P. Rosso and A. Lyhyaoui, "Implementation of the ArabiQA Question Answering System's Components," Proc. Of Workshop on Arabic Natural Language Processing, 2nd Information Communication Technologies Int. Symposium (ICTIS-2007), Fez, Morroco, pp. 3-5, April, 2007.

[42] B. A. Shawar, "A Chatbot As a Natural Web Interface to Arabic Web QA," International Journal of Emerging Technologies in Learning (iJET), vol. 6, pp. 37-43, 2011.

[43] O. Trigui, L. H. Belguith and P. Rosso, "Arabic Cooperative Answer Generation via Wikipedia Article Infoboxes," Research in Computing Science, vol. 132, pp. 129-153, 2017.

[44] N. Y. Habash, A. Soudi and T. Buckwalter, "On Arabic Transliteration," Arabic Computational Morphology: Knowledge-based and Empirical Methods, vol. 38, A. Soudi, A. v. d. Bosch and G. Neumann, Eds., Springer, pp. 15-22, 2007.

[45] Y. Benajiba, M. Diab and P. Rosso, "Using Language -Independent and Language- Specific Features to Enhance Arabic Named Entity Recognition," The International Arab Journal of Information Technology, vol. 6, 2009.

[46] M. C. De Marneffe and C. D. Manning, Stanford Typed Dependencies Manual, Stanford University, Ed. 2008, pp. 338-345, Sep. 2008.

[47] R. Al-Shalabi, G. Kanaan, J. M. Jaam, A. Hasnah and E. Hilat, "Stop-word Removal Algorithm for Arabic Language," Proceedings of International Conference on Information and Communication Technologies: From Theory to Applications, Damascus, Syria, 2004.

[48] H. Rodríguez, D. Farwell, J. Farreres, M. Bertran, M. Alkhalifa, M. A. Martí, W. Black, S. Elkateb, J. Kirk, P. Vossen and C. Fellbaum, "Arabic WordNet: Current State and Future Extensions," Proceedings of the 4th Global WordNet Conference (GWC 2008), Szeged, Hungary, 2008.

[49] Y. Regragui, L. Abouenour, F. Krieche, K. Bouzoubaa and P. Rosso, "Arabic WordNet: New Content and New Applications," Proceedings of the 8th Global WordNet Conference, pp. 330-338, 2016.

[50] Y. Benajiba, "Test-Bed for Passage Retrieval (PR) and Question Answering (QUA) Tasks," Y. Benajiba, Ed., [Accessed October 2017].

[51] B. Hammo, S. Abuleil, S. Lytinen and M. Evens: "Experimenting with a Question Answering System for the Arabic Language," Computers and the Humanities, vol. 38, no. 4, pp. 397-415, 2004.

ملخص البحث:

يتضمن الاهتمام المتزايد بمعالجة نصوص اللغة العربية وبحوث الشبكة العنكبوتية في مجال دلالات الألفاظ وتطورها حاجةً إلى تطوير أنظمة جديدة للإجابة عن الأسئلة. وتسمح هذه الأنظمة للمستخدمين بطرح أسئلة باللغة العربية والحصول على إجاباتٍ لها.

وتجدر الإشارة إلى أن غالبية أنظمة الإجابة عن الأسئلة المتاحة على الشبكة تركز على اللغة الإنجليزية واللغات ذات الأصل اللاتيني، مع القليل من الجهد الذي يركز على اللغة العربية التي تندرج تحت مسمّى اللغات الساميّة.

والبحث الذي بين أيدينا نسخة مبكرة من نظام جديد للإجابة عن الأسئلة باللغة العربية باستخدام البيانات المرتبطة. والنظام المقترح يهدف إلى مساعدة المستخدمين العرب في استكشاف دلالات الألفاظ وتطورها على الشبكة العنكبوتية.

في هذا البحث، نصف بتفصيلٍ كافٍ الوحدات التي يتألف منها نظام الإجابة عن الأسئلة المقترح لمعالجة الأسئلة المطروحة باللغة العربية لغويّاً من أجل إيجاد الإجابات الصحيحة لها. وقد أجريت تجارب لتقييم النظام المقترح وبيان فاعليته.