ISSN 1870-4069

Arabic Cooperative Answer Generation via Wikipedia Article Infoboxes

Omar Trigui[1], Lamia Hadrich Belguith[2], Paolo Rosso[3]

[1] ANLP Research Group, MIRACL Lab, ISAEG, University of Gafsa, Tunisia

[2] ANLP Research Group, MIRACL Lab, FSEG, University of Sfax, Tunisia

[3] PRHLT research center, Universitat Politècnica de València, Spain

omar.trigui@fsegs.rnu.tn, l.belguith@fsegs.rnu.tn, prosso@dsic.upv.es

Abstract. The typical question-answering system is facing many challenges related to the processing of questions and information resources in the extraction and generation of adequate answers. These challenges increase when the requested answer is cooperative and its language is Arabic. In this paper, we propose an original approach to generate cooperative answers for user-definitional questions designed to be integrated in a question-answering system. This approach is mainly based on the exploitation of the semi-structured Web knowledge which consists in using features derived from Wikipedia article infoboxes to generate cooperative answers. It is globally independent of a particular language, which gives it the ability to be integrated in any definitional question-answering system. We have chosen to integrate and experiment it in a definitional question-answering system dealing with the Arabic language entitled DefArabicQA. The results showed that this system has a significant impact on the approach efficiency regarding the improvement of the quality of the answer.

Keywords: Natural language processing, natural language generation, data extraction and integration, web knowledge, question answering system, cooperative answers, Arabic language.

1 Introduction

In this paper, we propose an approach that allows the generation of definitional cooperative answers using semi-structured Web knowledge bases. This approach is designed to be integrated in a definitional question-answering system. Before presenting the approach in detail, we will specify the general context, the motivations, the challenges to face, and the objectives to be achieved through this approach.

A question-answering system presents the intersection of several domains, such as information retrieval, information extraction, and natural language processing. Its typical interest is to make access to information toward information resources of large sizes and with heterogeneous, fast and smooth structures. A huge progress in results is achieved. This has been proved by organizing a series of competitive workshops dealing with the question-answering track by international conferences, such as TREC[1], CLEF[2] and NTCIR[3]. Different types of questions were dealt with these competitive workshops such as the complex one like "definition" question type. Typically, a question of the type "definition" is the one that asks about important information related to a fact, a person, an organization or an event. The adopted answer form to a definition question by competitive workshops is a list of information nuggets (i.e. a set of pieces of important informational texts) [Voorhees, 2003].

However, this answer form remains modest at the level of the structure and harmonization of information compared to what is expected by a user as an expected defining answer. This one presents for us a motivation to enhance this form of answer and to propose a new answer form reflecting high level coherence information and well-structured definition answer.

Fig. 1. An extract of a Wikipedia article entitled Mark Zuckrberg[4].

Many challenges arise when we get into dealing with the question-answering systems. We cite two aspects that are very important for their functioning. The first can handle information resources that are not well-formed, while the second one can find out the type of information looked up behind the question and deduce the details of the expected answer. These challenges increase when we deal with the "definition"

[1]http://trec.nist.gov/
[2]http://clef2015.clef-initiative.eu/
[3]http://research.nii.ac.jp/ntcir/index-en.html
[4]http://en.Wikipedia.org/wiki/Mark_Zuckerberg

question type and with a developed answer form for the expected defining answer adding the specificities of a language characterized by their low resources.

In this paper, we propose an approach which permits to generate a cooperative answer in the form of a paragraph designed to be integrated in a definition question-answering system dealing with low resource languages. The concept of this approach is based on the exploitation of semi-structured Web knowledge bases and specifically Wikipedia article infoboxes. We have chosen to exploit Wikipedia as it is a much large semi-structured Web knowledge base which contains more than 30 million articles in 287 languages. On the one hand, its website is the fifth most visited website in the world with 18 billion visitors[5], which can prove the confidence of their information and their wide coverage of topics, on the other hand. Among the components of a Wikipedia article, there is a one entitled infobox which contains a summary of important information relative to the main subject dealt with in a given Wikipedia article. These pieces of information are often located in a formatted box at the top of a Wikipedia article.

Figure 1 shows an example of an extract of a Wikipedia article entitled Mark Zuckrberg, while figure 2 shows an example of an infobox extracted from the Wikipedia article entitled "Jimmy Wales".

Fig. 2. An example of a Wikipedia article infobox entitled Jimmy Wales[6].

[5]http://en.Wikipedia.org/wiki/Wikipedia
[6]http://en.Wikipedia.org/wiki/Jimmy_Wales

2 State of the Art

In this section, we will present the main research studies based on the exploitation of Wikipedia as a Web knowledge base, as well as the major studies on generating answers in question-answering tasks. The exploitation of Wikipedia as a Web knowledge base has been introduced in various research studies addressing information retrieval, information extraction, construction of multilingual corpus and automatic translation [Lopez et al., 2011]. Among these research studies, we can mention those of Bizer et al. [2009] and Yahia & Salhi [2014], which are based on the exploitation of Wikipedia as a Web knowledge base in the information retrieval field. Bizer et al. [2009] made a great effort to extract structured information from Wikipedia and make it accessible to the Web. The resulting DBpedia knowledge base currently describes more than 2.6 million entities. However, Yahia & Salhi [2014] used Wikipedia as a knowledge base for the categorization of documents. Other research studies exploiting Wikipedia article infoboxes and addressing information retrieval appeared in a set of workshops. One of these interesting workshops is KBP[7] "Knowledge-Base Population" which has been organized by the TAC[8] conference since 2009 [Ji and Grishman, 2011; Surdeanu, 2013].

Regarding question-answering systems, further research studies based on Wikipedia have been designed. We can cite those which exploit Wikipedia as a knowledge base like [Trigui et al. 2010a; Brzeski&Boi´nski, 2014; Yang et al. 2014] and [Ryu et al. 2014]. The research of Trigui et al. [2010a] exploited Wikipedia article contents through a Web search engine in an Arabic definition question answering system to build a specific information resource relative to each given question. On the other hand, Yang et al. [2014] proposed a method to build a robust knowledge resource based only on semantic associations automatically extracted from Wikipedia. The obtained knowledge resource was designed to be integrated in a question-answering system. On their part, Breski & Boinski [2014] proposed a method which is based on associations between Wikipedia articles to answer factual questions. Moreover, Ryu et al. [2014] proposed a method to categorize the Wikipedia structures into article contents, infoboxes, category structures, article structures and redirection links. These Wikipedia structure categories were designed to be used as a rich knowledge resource for factual question-answering systems. Other research studies dealing with question-answering systems exploited Wikipedia for the validation of answers [Buscaldi & Rosso, 2006; Cui et al. 2007].

We have cited examples of research studies based on Wikipedia exploitation as a Web knowledge base in the domains information retrieval and question-answering. Here, we cover the details of the major studies on generating answers in question-answering. Many research studies have dealt with the question-answering systems, but only a few of them have addressed the answer-generation step beyond the answer extraction step [Voorhees, 2004; Dang et al., 2007]. Typically, the answer-generating step permits to generate an answer where there is more than one possible answer or no

[7]http://pmcnamee.net/kbp.html
[8]http://www.nist.gov/tac/tracks/index.html

answer found in the data resources [Benamara & Saint-Dizier, 2004]. It would be an indirect answer to the user's question and more helpful than the direct one [Corella & Lewison, 2009]. This form of answer is entitled 'a cooperative answer' [Benamara, 2004]. Most of the research studies dealing with cooperative answer generation are based on integrating knowledge representation and reasoning mechanisms. We can mention, for example [Prager et al. 2003; Benamara & Saint-Dizier, 2004] and [Lupkowski & Leszczyńska-Jasion, 2014]. Prager et al. [2003] proposed to answer a definition question by gathering answers to factual questions derived from the given definitional question. As for Benamara & Saint-Dizier [2004], they proposed an approach that enables to answer factual questions in French by cooperative answers dealing with the tourism domain. On their part, Lupkowski & Leszczyńska-Jasion [2014] described a system designed to generate cooperative answers based on inferential erotetic logic concepts.

We will focus now on the study of research studies addressing the Arabic Web knowledge base in information retrieval and question-answering fields. In the literature, research studies dealing with the Web knowledge base in information retrieval and addressing the Arabic language have witnessed a growing interest during the last few years [Ezzeldin&Shaheen, 2012; AlZoghbyaa et al., 2013]. We can mention in this respect the research studies of Beseiso et al. [2011], Al-Zoghby and Shaalan [2015], and Al-Bukhitan et al. [2014]. These research studies have shared the goal of facilitating the search and the access to information adopting the Semantic Web technology in information retrieval.

Beseiso et al. [2011] proposed a new framework intended to add a semantic Web layer to the current Web-based applications in order to improve the searching and linking processes. However, Al-Bukhitan et al. [2014] proposed an automatic annotation tool that supports the semantic annotation of Arabic Web documents for semantic search engines. A promising performance was achieved by this automatic annotation tool. On the other hand, Al-Zoghby and Shaalan [2015] proposed a semantic search approach applied to Arabic Web content which is based on the Vector Space Model. It consists in locating Web contents that are semantically related to the query's concepts rather than relying on the exact matching with keywords in queries.

In spite of the efforts made in the Arabic language to adopt Web knowledge in information retrieval, there is a lack of research studies adopting and exploiting Web knowledge in question-answering. This can be explained by the structure of the Arabic Web knowledge bases which make it particularly difficult to handle the automatic processing challenges of the Arabic language properties by question-answering systems. We mention two research studies which used the Web as a knowledge base [Trigui et al, 2010a; Hasanain et al. 2014]. The first is based on Web search engines to construct specific resource knowledge of snippets for each given question in a definition question-answering system. The second is based on Twitter to construct corpus constituted of millions of tweets for its system [Hasanain et al., 2014].

Apart from that, there are various research studies dealing with Arabic question-answering systems based on closed corpus of documents [Ezzeldin & Shaheen, 2012; Shaheen & Ezzeldin, 2014]. With respect to generating answers in the Arabic language, we have to mention that, to our knowledge, there are practically no research

studies that addressed this issue. Most prominent Arabic question-answering systems return the answers in a paragraph form, such as Hammo et al. [2004] or in a list of information nuggets, as in the case of Trigui. [2011]; Badawy et al. [2011]; Fareed et al. [2014] and Kurdi et al. [2014].

To sum up, we have cited research studies dealing with Wikipedia as a Web knowledge base in various manners for question-answering systems. Among these research studies, we have mentioned question-answering systems which generate cooperative answers based on integrating knowledge representation and reasoning mechanisms. In addition, we have shown the important research studies addressing Arabic question-answering systems and their various forms of answers. As for our approach, it is meant to exploit the wealth of Wikipedia information, and especially Wikipedia article infoboxes, for generating answers to definition questions in the context of a definition question answering system.

3 Approach for Generating Cooperative Answers

In this section, we detail the proposed approach to generate cooperative answers to definitional questions for question-answering systems. It is based on a part of the semi-structured Web knowledge base presented by the Wikipedia article infoboxes. It can be integrated in each question-answering system dealing with definitional questions, independently of its particular language. This approach involves three main tasks: the infobox class generation, the cooperative answer pattern generation and the cooperative definition answer generation (see figure 3). Hereafter, we will detail these three tasks.

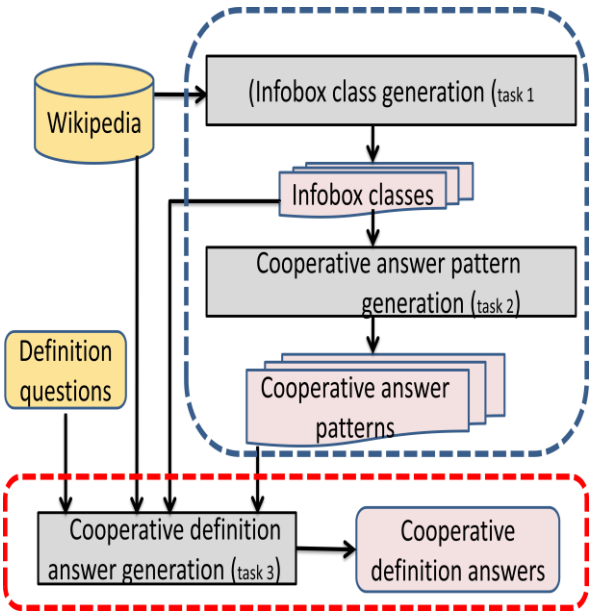

Fig. 3. Main tasks of the proposed approach.

3.1 Infobox Class Generation (Task 1)

The infobox class generation task includes a set of three sub-tasks, namely: Wikipedia article selection, infobox exploitation and infobox grouping, as mentioned in figure 4. These sub-tasks are based on the information in the Wikipedia article infoboxes. It exploits the hypothesis that a given Wikipedia article infobox shows a resource of specific and relevant information relative to a definite named entity in a given language.

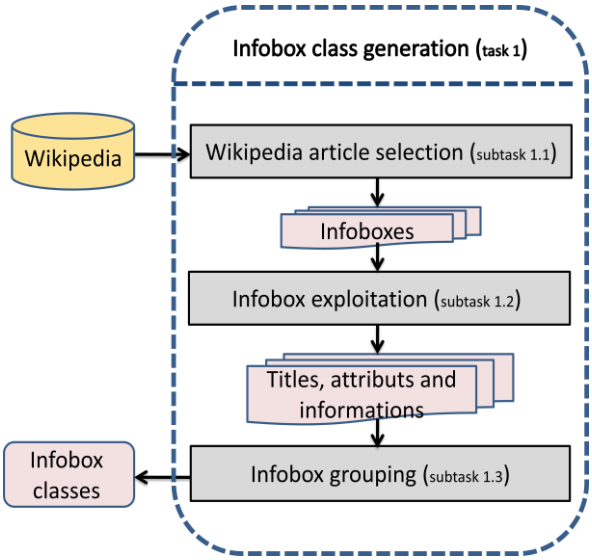

Fig. 4. The subtasks of the infobox class generation task.

Wikipedia Article Selection (sub-task 1.1). Wikipedia contains hundreds of thousands of articles in different languages where each article represents a separate Web page describing a definite named entity (e.g. an event, a person, an organization or a concept). This sub-task consists of browsing Wikipedia for a given language in order to select Wikipedia articles containing infoboxes. The number of Wikipedia articles varies from one language to another. It reaches more than four million items in the English language and around three hundred thousand items in the Arabic language, as an example[9].

Infobox Exploitation (sub-task 1.2). This sub-task consists in extracting infobox features (i.e. attributes, title and information) from Wikipedia articles containing Infoboxes. Figures 2 and 5 present two infoboxes containing summary of specific information respectively describing Jimmy Walles and Bill Gates (two Wikipedia article titles). Table 1 shows attributes extracted from two infoboxes, respectively, enti-

[9]http://en.wikipedia.org/wiki/Wikipedia#Language_editions

tled “Jimmy Wales” and “Bill Gates”. The first infobox attributes are “born”, “residence”, “other names”, “alma mater”, “occupation”, “title”, “successor”, “board member of”, “spouse”, “awards” and “Website” while the attributes of the second one are “born”, “residence”, “alma mater”, “occupation”, “active years”, “net worth”, “board member of”, “spouse”, “children”, “parents” and “Website”. These examples confirm that each infobox presents a resource of the basic information describing a precise named entity which can be a person, an organization, a date, a location, or an event. In our case, infobox attributes present the most important features of an infobox.

Fig. 5. Example of an infobox entitled Bill Gates[10].

Table 1. Attribute list extracted from the infoboxes entitled “Jimmy Wales” and “Bill Gates”.

Jimmy Wales infobox attributes		**Bill Gates infobox attributes**	
Born	***boardmember of***	***Born***	***boardmember of***
Residence	*successor*	***Residence***	***Spouse***
other names	***spouse***	***alma mater***	*Children*
alma mater	*awards*	***occupation***	*Parents*
Occupation		*active years*	*Signature*
Title	***website***	*net worth*	***Website***

[10]http://en.wikipedia.org/wiki/Bill_Gates

Grouping Infoboxes (subtask 1.3). The sub-task consists in grouping infoboxes in a class of infoboxes according to a minimum rate of similarity between their attributes (those in common). For each infobox class, we consider the following features: a label (i.e. a semantic category), which is typically a hyperonymy of the included infoboxes in this class, as well as a set of attributes shared by the various semantically closest infoboxes.

Table 2. The infobox class features entitled "entrepreneur".

Semantic category	
Entrepreneur	رجل الأعمال
Attributes	
Born	ولد/ت
Residence	إقامة
Alma mater	جامعة
Occupation	ت/يشغل منصب
Board member of	عضو مجلس إدارة
Spouse	شريك حياته/ها
Website	موقع الويب

Table 3. The infobox class features entitled "political party".

Semantic category	
Political party	حزب سياسي
Attributes	
Founded by	تأسس من طرف
Foundation year	سنة التأسيس
party leaders	قادة الحزب
Ideology	الإيديولوجية المتبعة
Location	الموقع
the leader	القائد
number of deputies	عدد النواب
leadership center	مركز القيادة
Website	موقع الويب

The result is of this task is a set of infobox classes where each infobox class contains the semantically closest infoboxes from the set of Wikipedia article infoboxes. A semantic category is attributed to each infobox class by a person who would be a reader of the international newspapers and native speaker of the respective language of the infoboxes.

Typically, these semantic categories provide titles for infobox classes. Tables 2, 3 and 4 show three infobox classes with their respective features mentioned in English and in Arabic. For example, the infobox class cited in table 2, which has the semantic category 'entrepreneur', is the result of the grouping of the semantically closest infoboxes, such as the infoboxes entitled Jimmy Wales and Bill Gates relatively in figures 2 and 5. Their shared attributes, which are mentioned in bold in table 1, constitute the attributes of this infobox class, such as "born", "residence", "alma mater", "occupation", "board member of", "spouse" and "Website".

Table 4. The infobox class features entitled "sports team".

Semantic category	
Sports team	فريق رياضي
Attributes	
full name	الاسم الكامل
Nickname	الكنية
Founded	أسّس
Stadium	الملعب
League	الدوري
Coach	المدرب
Website	موقع الويب

3.2 Cooperative Answer Pattern Generation (Task 2)

The cooperative answer pattern generation task consists in building a set of cooperative answer patterns where each one is taken as a skeleton of a definition answer. It consists in exploiting infobox class features. The sub-tasks are involved in this task are, respectively, composing skeletons and formulation of cooperative patterns (see figure 6).

Composing skeletons (sub-task 2.1). This subtask consists in generating a pattern answer skeleton for each infobox class using all its attributes. The attributes of each infobox class are taken in order, one by one, in the composition of the respective pattern answer skeleton in the first step. In the second step, a blank reserved for respective information is added after each attribute in the composing pattern answer skeleton.

Formulation of cooperative answer patterns (sub-task 2.2). This sub-task consists in checking each composed pattern answer skeleton and adding specific punctuation marks (i.e. a comma, a full stop, etc.) after each blank reserved for information in the patterns. The choice of the respective punctuation marks is made by a linguistic expert.

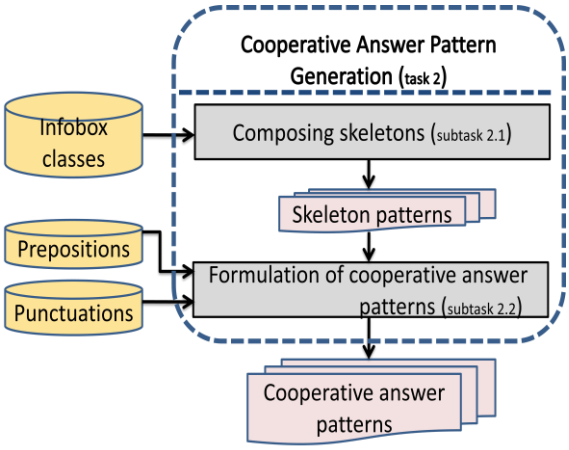

Fig. 6. Subtasks constituting the cooperative answer pattern generation task.

Table 5. The cooperative answer pattern relative to the semantic category "entrepreneur".

A COOPERATIVE ANSWER PATTERN
<VALUEQUESTION FOCUS> + IS +{BORN}+ IN + <VALUE> + . + HIS {RESIDENCE}+ IS AT + <VALUE> + . + HIS {ALMA MATER }+ IS AT THE + <VALUE> +.+ HIS + {OCCUPATION}+ IS/ARE + <VALUE> +,+ {BOARD MEMBER OF} + <VALUE> +. +HIS + {SPOUSE} + IS + <VALUE>+. +HIS + {WEBSITE} + IS + <VALUE>+.
<المعلومةموضوع سؤال > {ولد/ت} + في + > المعلومة> +. + {تقيم/يقيم} + في+<المعلومة> +.+ {متخرج/ة من } + <المعلومة> +.+ {يشغل منصب} + <المعلومة> +،+ { عضو مجلس إدارة } + <المعلومة> +. + {شريك حياته/ها} + هو/هي + <المعلومة> +.+{ موقع الويب الخاص+}هو+<المعلومة>.+

Table 6. A cooperative answer pattern relative to the semantic category "political party".

A COOPERATIVE ANSWER PATTERN
<VALUE QUESTION FOCUS> + IS + {FOUNDED BY} + <VALUE>+ . + THE + {FOUNDATION YEAR} + IS + <VALUE>+ . + THE + {PARTY LEADERS}+ ARE + <VALUE> + . + IT + {IDEOLOGY} + <VALUE> +.+ THE {LOCATION} + IS + <VALUE> +. THE + {LEADER}+ IS + <VALUE> +. THE + {NUMBER OF DEPUTIES} + IS + <VALUE> + AND THE + {LEADERSHIP CENTER}+ IS + <VALUES> +.
<المعلومةموضوع سؤال > {تأسس من طرف } + <المعلومة> +. +{سنة التأسيس} + كانت في+ <المعلومة> +. { قادة الحزب } + هم+<المعلومة> +. + {الإيديولوجية المتبعة} + هي +<المعلومة> +.+ { الموقع} + هو +<المعلومة> +.+ {القائد } + هو/هي +<المعلومة> +.+ { عدد النواب } + هو+ <المعلومة> +.+ { مركز القيادة } + هو+ <المعلومة> +.

These two sub-tasks enable to generate a cooperative answer pattern for each infobox class. Tables 5 and 6 show two examples of cooperative answer patterns in English and in Arabic. The first is related to the infobox class having the semantic

category "entrepreneur", while the second is related to the infobox class having the semantic category “political party”. The texts inside the curly braces are these attributes while the annotation "<value>" means the respective information of the associated attribute. The annotation “<value $_{\text{Focus question}}$>” means the focus of the given question.

3.3 Generating Answers (Task 3)

The answer generation task is composed of three sub-tasks. It consists in generating cooperative definition answers to definition questions. The first sub-task consists in extracting the main named entity of the given definitional question (i.e. the question focus). The second one consists in selecting the adequate cooperative answer pattern relative to the given question, while the last sub-task consists in filling the blank of the selected answer pattern with the respective information. We will detail these sub-tasks one by one below.

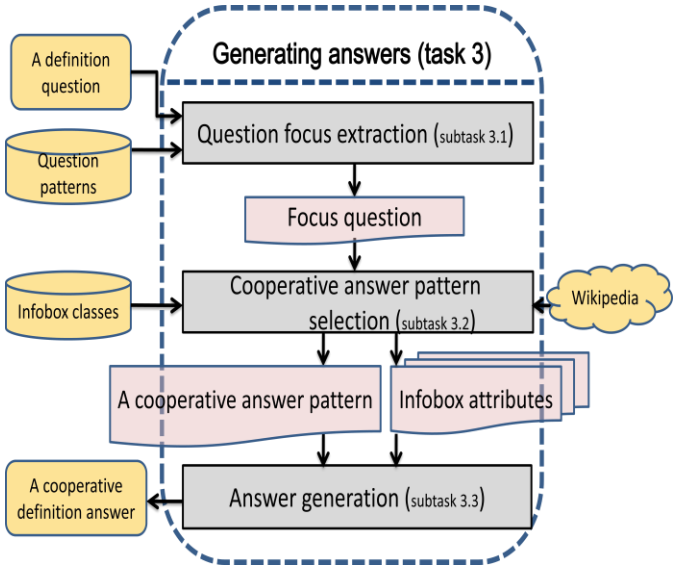

Fig. 7. Sub-tasks constituting the generating answer task.

Question Focus Extraction (subtask 3.1). It consists in identifying the main named entity presenting the interest subject of a definition question (i.e. the question focus). The identification is based on a set of lexical patterns designed for definition questions. Table 7 shows examples of typical lexical patterns to identify question focuses from definition questions in English and in Arabic.

Table 8 presents an example of a definition question as well as its respective lexical pattern of definition questions in English and in Arabic. The definition question is «Who is Steve Chen? » and its respective question focus is « Steve Chen » identified by the following lexical pattern: “Who + be + <a question focus>+ ?”.

Table 7. Lexical patterns of the definition questions in English and in Arabic [Benajiba et al., 2014].

Definition question patterns	Expected answer types
Who+be+<a question focus>+ ? من هو	من هي <الموضوع>+؟	*Interesting information about a person*
What+be+<a question focus>+? ما هو	ما هي<الموضوع>+؟	*Interesting information about an organization or a concept*

Table 8. An example of identifying a question focus.

Definition question	Lexical pattern	Named entity
Who is Steve Chen ?	*Who+be+<a question topic>+ ?*	*Steve Chen*
من هو ستيف تشين ؟	***من هو+ <الموضوع>+؟***	ستيف تشين

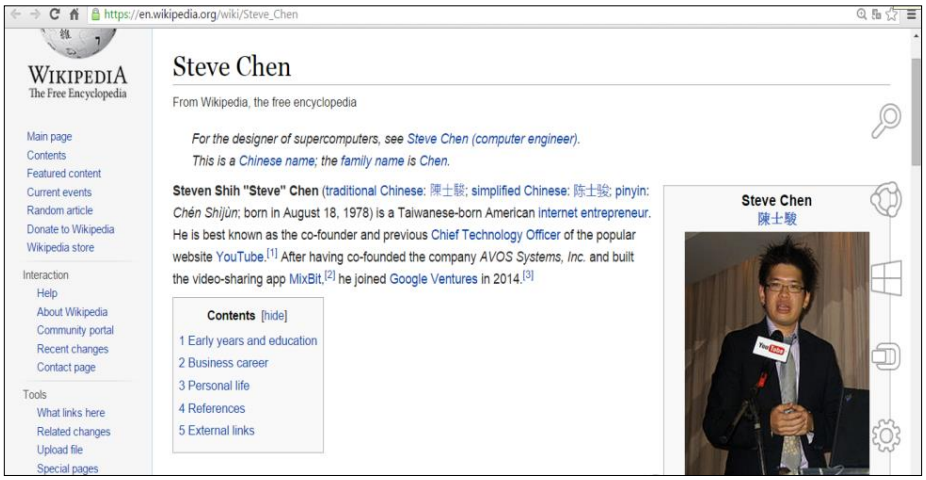

Fig. 8. An extract of a Wikipedia article entitled "Steve Chen".

Cooperative Answer Pattern Selection (sub-task 3.2). Two steps constitute this sub-task. The first consists in looking for the Wikipedia article infobox relative to the extracted question focus in the previous subtask. In case there is an infobox that which has a title identical to the question focus, the Wikipedia article infobox features are extracted and compared to the attributes of each infobox generated in task 1. The

second step consists in looking for the most suitable infobox class of the selected Wikipedia article infobox.

Fig.9. A Wikipedia article infobox entitled "Steve Chen" with its attributes.

A process of identifying the highest overlap rate between their attributes is lunched. The suitable infobox class selected is the one having the highest overlap rate of attributes. Its cooperative answer pattern is chosen as the adequate skeleton of the expected definition of the cooperative answer to generate. We take as an example the question cited in table 8 to show how to choose its suitable cooperative answer pattern. Figure 8 shows an extract of the Wikipedia article entitled "Steve Chen" (i.e. the identified question focus). To identify the adequate cooperative answer pattern, we compare the extracted infobox attributes (figure 9) to the attributes characterizing each generated infobox class. In this case, the cooperative answer pattern selected for the current question is the one assigned to the infobox class having as semantic category "entrepreneur" (see table 5).

Table 9. An example of a cooperative definition answer with its respective cooperative answer pattern.

<VALUE QUESTION FOCUS> + WAS +{BORN}+ IN + <VALUE> + . + HIS {RESIDENCE}+ IS +AT + <VALUE> + . + HIS +{ALMA MATER }+ IS +IN+ <VALUE> +.+ HIS + {OCCUPATION}+ IS/ARE + <VALUE> + , + {BOARD MEMBER OF} + <VALUE> +.+ HIS + {SPOUSE} + IS + <VALUE> +.+HIS + {WEBSITE} + IS + <VALUE> +.
Steve Chen was born in Taipei, Taiwan. His residence is in San Francisco, California, US. His alma mater is at the University of Illinois at Urbana-Champaign. His occupation is co-founder of Avos systems. His spouse is Park ji-hyun (Jamie Chen).
<المعلومة موضوع سؤال > {ولد/ت} + في + < المعلومة > +. + {تقيم/يقيم} + في+ <المعلومة> +.+ {متخرج/ة من} +<المعلومة> +.+ {يشغل منصب} + <المعلومة> +. + {شريك حياته/ها} + هو/هي + <المعلومة> +.+{ موقع الويب الخاص+}+هو+<المعلومة>+.
ستيف تشين ولد في تايبيه، تايوان. يقيم فيسان فرانسيسكو، كاليفورنيا، الولايات المتحدة. متخرج من جامعة إلينوي في أوربانا شامبين. يشغل منصب مؤسس مشارك ل"أفوس سيستم" . شريك حياته بارك جي-هيون (جيمي شين).

Answer Generation (subtask 3.3). It consists in generating definition answers by filling the blanks of the selected cooperative answer patterns by taking into consideration the correspondence between the cooperative answer pattern attributes and the extracted Wikipedia article infobox attributes. The generated answer is characterized by its cooperative form. For the question cited in table 8 "Who is Steve Chen?", the selected cooperative answer pattern is filled with the respective information extracted from the Wikipedia article infobox entitled "Steve Chen" (see figure 9). Each attribute of the cooperative answer pattern does not have its respective information is removed, while the other ones are kept. Table 9 shows the selected cooperative answer pattern and the generated cooperative answer in English and also with its translation in Arabic.

4 Answer Generation Module towards a Question Answering System

In order to evaluate the impact of the proposed approach for generating cooperative answers and to facilitate its integration in a definition question-answering system, we have implemented it in a module entitled cooperative answer generation. Two processing phases are required to realize this implementation of the proposed approach in a module: "An off-line processing phase" and "an on-line processing phase". The former includes common processing which is not associated with a given definition question (see figure 9) while the latter deals with treatments specific to a given defini-

tion question (see figure 11). As our approach can be applied to any language, we have decided to choose a low resource language, such as Arabic. This choice gives us more chance to deduce the limits and the performance in the difficult cases. The selected language is used for the resource, the questions and the answers.

The Off-line Processing Phase. This phase includes a series of treatments divided into two steps to exploit Wikipedia information. These steps are respectively "infobox class generation" and "cooperative answer pattern generation" (see figure 10).

Infobox Class Generation. This step consists in collecting Wikipedia articles, extracting the infoboxes, then grouping them in classes. The collection process consists in collecting all the Arabic Wikipedia articles from the website of Wikipedia[11]. The collection process is realized automatically through a tool entitled WikiPageDownload, developed by our research team. To a list of named entities, which exhibits the Wikipedia article titles in a given language, permits to download and save the content of each Wikipedia article associated with each named entity in the given list. For the Arabic language, 321454 Wikipedia articles are downloaded and saved. Figure 10 shows a distribution of these Wikipedia articles according to the existence of infoboxes, while figure 11 presents some examples of Wikipedia article titles.

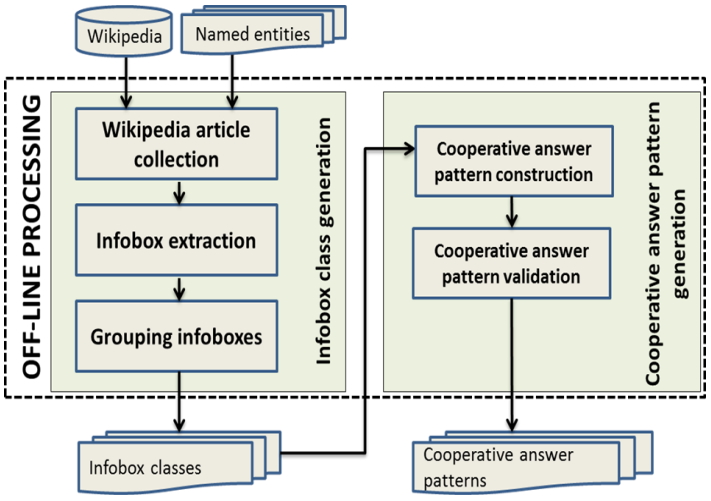

Fig. 10. The implementation process of the off-line phase.

The extraction of infoboxes process consists in extracting infobox features of 78760 downloaded Wikipedia article infoboxes (see Table 10). The process of grouping infoboxes consists in exploiting the downloaded infobox features by clustering together the infoboxes having highest overlap rate between their attributes. From

[11]https://sites.google.com/site/omartrigui/downloads

78760 Wikipedia article infoboxes, 189 infobox classes were generated; each infobox class contains nearly 417 infoboxes (see table 11).

Table 10. Arabic Wikipedia article distribution.

Wikipedia articles containing *infobo*xes	78760	**24,50%**
Wikipedia articles containing no infoboxes	242694	75,50%
Total Wikipedia articles in the Arabic language	321454	100%

Table 11. The average infoboxes per infobox class.

Number of infobox classes	189
Average infoboxes per infobox class	417

For each infobox class, a semantic category is attributed by a human expert presenting a hyperonymy of all their respective infobox titles. Tables 2, 3 and 4 present respectively three infobox classes respectively characterized by their attributes and their semantic categories, which are given hereafter: “entrepreneur”, “political party” and “sports team”.

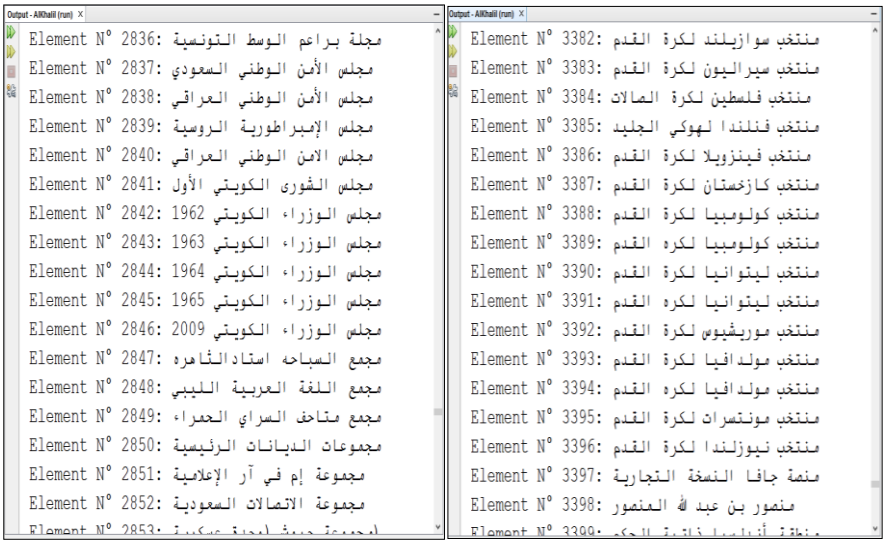

Fig. 11. An extract of the Wikipedia article titles list in the Arabic language.

Cooperative answer pattern generation. It consists in generating a cooperative answer pattern using infobox class attributes. For each infobox class, the respective attributes are taken in their order of appearance and followed by blanks and specific

punctuation marks (i.e. commas, full stops, etc.). This step enables to associate a cooperative answer pattern for each infobox class. For 189 infobox classes, we obtained 189 cooperative answer patterns.

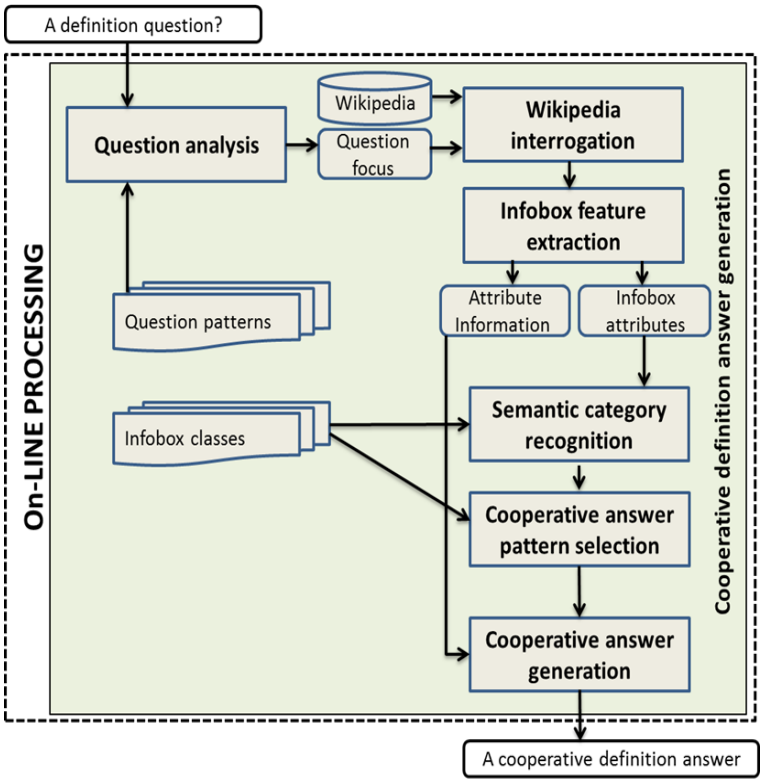

Fig. 12. The implementation process of the on-line phase.

The On-line Processing Phase. The on-line processing phase is based on the results of the off-line processing phase to generate cooperative definition answers. It is constituted by six processing steps: question analysis, Wikipedia interrogation, infobox feature extraction, semantic category recognition, cooperative answer pattern selection and cooperative answer generation (see figure 9). They are running for each given definition question to obtain the suitable cooperative definition answer. A cooperative definition answer does not exist in any documents but generated towards piece of information and a cooperative answer pattern.

5 Experiments

We will present the details of the experiments carried out to test the validity of the cooperative answer generation module and evaluate its impact. We have selected

DefArabicQA system (i.e. a definition question-answering system dealing with the Arabic language) to integrate this module and realize experiments with it [Trigui et al, 2010a]. DefArabicQA is a system based on both linguistic and frequency-based approaches. It uses a surface pattern technique to extract candidate answers and statistical features to rank them. It is based on Web search engines as knowledge bases [Trigui et al, 2010b]. Its architecture is illustrated in figure 13.

Experimental Data. Before illustrating the experimental results, we will describe the test data and the used performance measure. We have used a dataset comprising 300 definition questions in the Arabic language. The questioner is an adult, a native speaker of Arabic, and a reader of Arabic newspapers. Table 12 shows a part of these questions.

To measure the performance of the carried out experiments, the accuracy measure is used typically to evaluate the overall quality of a question answering system providing one potential answer for a given question. It is a number between 0 and 1, which indicates the probability of a question answering system to provide the correct answer on average. It is expressed as follows: *Accuracy = Number of correct answers / Number of questions*.

Evaluation Methodology. Three experiments are carried out. The first experiment is executed using the Google Web search engine (a baseline); the second one is realized using DefArabicQA system and the third experiment is performed using the DefArabicQA system extended by the cooperative answer generation module (see figure 14). All the experiment results are compared to a baseline.

The accuracy of the integrated module must be deduced. All the experiments are carried out with the same question dataset. We have to note that two assessors have evaluated the returned answers for each run. Both of them are Arabic native speakers and Arabic newspaper readers. To count the correct answers of these experiments with a fairer measure, we took the following hypothesis: for the first experiment (i.e. using the Google Web search engine), a question can be answered only if the first top snippet returned by the Google Web search engine containing at least one information nugget. For the second experiment with the DefArabicQA system, a question is annotated answered correctly only if its corresponding answer contains at least one information nugget without extraneous information. Then, for the third experiment with the DefArabicQA system extended with the cooperative answer generation module, a question can be answered correctly only if its answer contains a cooperative definition answer or at least one information nugget without extraneous information.

Experimental Results. We will now deal with the experimental results of the carried out experiments, which are presented in Table 13. Regarding the first experiment, 45% of the questions were answered by the Google Web search engine from the first top snippet but the search engine failed to return the correct answer from the first top snippet to the rest (55% of the total questions). This experiment obtained 0.45 as an accuracy measure. It was taken as a baseline to the other two experiments.

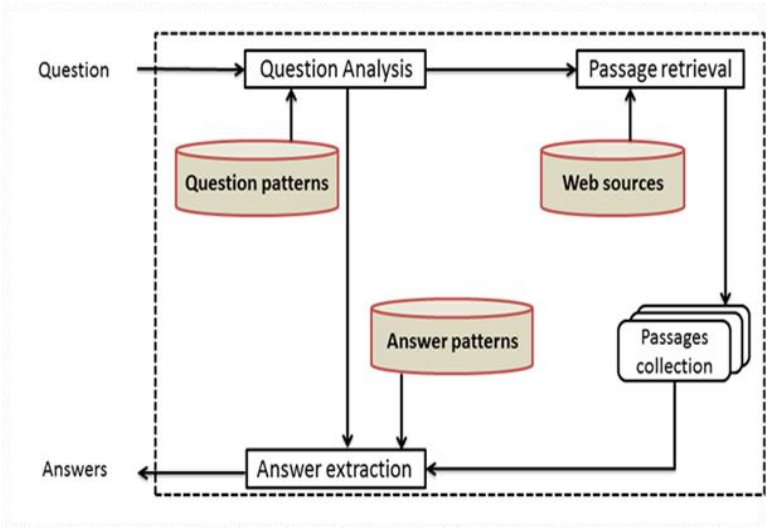

Fig. 13. DefArabicQA system architecture [Trigui et al, 2010b].

Table 12. A part of the definition questions of the test data.

Question N° 1 :	ما هو الاتحاد العام التونسي للشغل؟
	What is the Tunisian General Labor Union?
Question N° 2 :	ماهي الشركة السعودية للكهرباء؟
	What is the Saudi Electricity Company?
Question N° 3 :	ماهي شركة الزامل للاستثمار الصناعي؟
	What is the Zamit Industrial Investment Company?
Question N° 4 :	ماهي الشركة العامة للبريد والاتصالات السلكية واللاسلكية؟
	What is the General Post and Telecommunications Company?
Question N° 5 :	ماهي الشركة العربية للاستثمار؟
	What is the Arab Investment Company?
Question N° 6 :	ماهي الشركة العربية للعود؟
	What is the Arabian Oud Company?
Question N° 7 :	ماهي الشركة القابضة للنقل البحري و البري؟

	What is the Holding Company for Maritime and Land transport?
Question N° 8 :	ماهي الشركة العالمية للكتاب؟
	What is the World Book Publishing?

The results of the second experiment brought an overall improvement over the baseline. The DefArabicQA system succeeded in answering around 63%, but failed to answer around 37% of the questions. Accuracy is around 0.63, which it is over the baseline by 0.18.

In the third experiment, we tested whether the integration of the cooperative answer generation module can further improve the accuracy of the DefArabicQA system further. This experiment was carried out using DefArabicQA system extended by the cooperative answer generation module. A first part equal to 70% of the question set was answered; however, the remaining 30% of the questions were not answered in this experiment (see Table 13). The evaluation results obtained brought an improvement of 0.07 in the accuracy over the second experiment and 0.25 over the baseline.

Table 13. Experimental results.

	Answered questions (%)	Unanswered questions (%)
Google Web Search engine (baseline)	45	55
DefArabicQA system	63	37
Extended DefArabicQA system	70	30

5.1 Discussion

The results of the second experiment brought an overall improvement over the baseline. The DefArabicQA system succeeded in answering around 63%, which is important compared to the overall results released in TREC [Voorhees, 2003]. This fact explains why when we carried out the third experiment (i.e. using the DefArabicQA system with the cooperative answer generation module); we obtained an improvement of only 0.07% in the accuracy compared to the accuracy achieved in the second experiment (see figure 12 and Table 13). Indeed, a rise of the accuracy measure brought by the cooperative answer generation module, a significant improvement of the quality by more than 60% of the returned answers, is noticed. These answers have as a common feature their cooperative form; however, in the second experiment, they have been answered via the information nugget form.

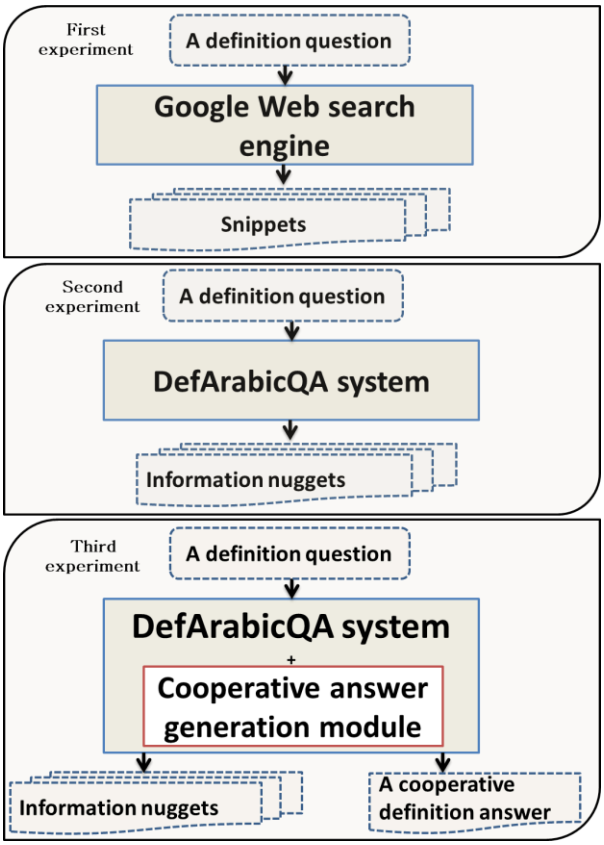

Fig. 14. The three carried out experiment architecture systems.

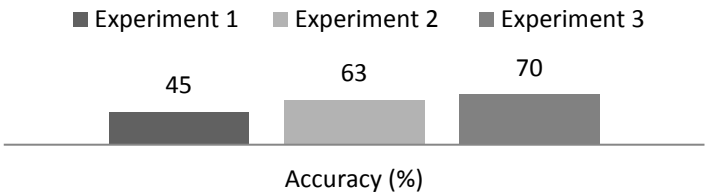

Fig. 15. Accuracy of the performed experiments.

We deduced from the percentage of questions answered by a cooperative answer that the cooperative answer generation module has succeeded in answering only from Wikipedia article infoboxes, up to 40% of the total test questions. This result is improved in spite of using only Wikipedia article infoboxes as a resource, knowing that the number of Arabic Wikipedia article infoboxes is very small compared to other languages. There are only around three hundred and twenty Wikipedia articles in Arabic compared to nearly four million Wikipedia articles in English. For the proposed module, the more Wikipedia articles there are, the more infobox classes we

have and the more successful we will be to generate a cooperative definition answer to a given definition question.

Thus, the fact that a portion of 60% of test questions which did not answered by a cooperative definition answer is caused mainly by the relative lack of Wikipedia articles in the Arabic language. In order to be more capable of dealing with this gap between languages, we try to look for information from Wikipedia article contents and not only from Wikipedia article infoboxes for certain languages. In general, the integration of the cooperative answer generation module in the DefArabicQA system has had a significant impact on the overall accuracy and on the quality of the returned answers. Therefore, it can be considered as a complementary module which has a positive influence on any definition question-answering system.

6 Conclusions and Future Work

We have proposed an approach for the generation of cooperative answers to definition questions. This approach is based on Wikipedia article infoboxes as a Web knowledge base. It is characterized by being language-independent and having the possibility to deal with open field questions. Its advantage is that it uses what is available on the Web to reach a cooperative definition answer, which should be consistent and informative, especially for a low resource language. Our experimental results show that the integration of the proposed approach in a definitional question-answering system dealing with the Arabic language has significantly outperformed the baseline which is based on Web search engines. In particular, we have shown that the Wikipedia article infoboxes can be used as a resource for generating definition cooperative answers. The limitations of the proposed approach are mainly related to the language adopted by the information resources.

As perspectives and in order to raise the effectiveness of this approach even if the specific Wikipedia articles do not contain infoboxes, we plan to exploit more Wikipedia article contents and test the validity of the approach in a multilingual context.

Acknowledgements. The work of the third author was partially funded by the Spanish Ministry of Economy, Industry and Competitiveness (MINECO) under the SomEMBED research project (TIN2015-71147-C2-1-P) and by the Generalitat Valenciana under the grant ALMAMATER (PrometeoII/2014/030).

References

1. Al-Bukhitan, S., Helmy, T., Al-Mulhem, M.: Semantic Annotation Tool for Annotating Arabic Web Documents. In: Proceedings of the 5th International Conference on Ambient Systems, Networks and Technologies (ANT-2014), vol. 32, pp. 429–436 (2014)
2. Al-Zoghbyaa, A. M., SharafEldin, A., Hamza, T. T.: Arabic Semantic Web Applications – A Survey. Emerging Technologies in Web Intelligence, vol. 5, no. 1, pp. 52–69 (2013)
3. Al-Zoghbyaa, A. M., Shaalan, K.: Conceptual Search for Arabic Web Content. In: Proceedings of the CICLing 2015, Part II, LNCS 9042, pp. 405–416 (2015)

4. Badawy, O., Shaheen, M., Hamadene, A.: ARQA High-Performance Arabic Question Answering System. In: Proceedings of Arabic Language Technology International Conference (ALTIC 2011), Alexandria, Egypt, pp. 129–136 (2011)
5. Benajiba, Y., Rosso, P., Abouenour, L., Trigui, O., Bouzoubaa, K., Belguith, H. L.: Question Answering. In: Natural Language Processing of Semitic Languages, pp. 335–370 (2014)
6. Benamara, F.: WebCoop: Un système de réponses coopératives sur le Web. Thèse de doctorat, Université Paul Sabatier (2004)
7. Benamara, F., Saint-Dizier, P.: Lexicalisation Strategies in Cooperative Question-Answering Systems. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING), Geneva, Switzerland, pp. 345–352 (2004)
8. Beseiso, M., Ahmad, A. R., Ismail, R.: An Arabic Language Framework for Semantic Web. In: Proceedings of the Semantic Technology and Information Retrieval (STAIR), Putrajaya, pp. 7–11 (2011)
9. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - A Crystallization Point for the Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 7, Issue 3, pp. 154–165 (2009)
10. Brzeski, A., Boinski, T.: Relation-based Wikipedia Search System for Factoid Questions Answering. International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 9, pp. 5601–5605 (2014)
11. Buscaldi, D., Rosso, P.: Mining knowledge from Wikipedia for the Question Answering task. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Genova, Italy, pp. 727–730 (2006)
12. Corella, F., Lewison, K. P.: A Brief overview of cooperative Answering. http://pomcor.com/whitepapers/cooperative_responses.pdf (2009)
13. Cui, H., Kan, M. Y., Chua, T. S.: Soft Pattern Matching Models for Definitional Question Answering. ACM Transactions on Information Systems (TOIS), Vol. 25, Issue 2 (2007)
14. Dang, H. T., Kelly, D., Lin, J.: Overview of the TREC 2007 question answering track. In: Proceedings of the 16th Text REtrieval Conference (TREC 2007), Maryland, USA (2007)
15. Ezzeldin, A. M., Shaheen, M.: A Survey of Arabic question answering: challenges, tasks, approaches, tools, and future trends. In: Proceedings of the 13th International Arab Conference on Information Technology (ACIT 2012), Jordan, pp. 280–287 (2012)
16. Fareed, N., Mousa, H., Elsisi, A.: Syntactic Open Domain Arabic Question/Answering System for Factoid Questions. In: Proceedings of the 9th International Conference on Informatics and Systems (INFOS2014), Natural Language Processing and Knowledge Mining Track, pp. 1–9 (2014)
17. Hammo, B., Ableil, S., Lytinen, S., Evens, M.: Experimenting with a Question Answering system for the Arabic language. Computers and the Humanities, Vol. 38, Issue 4, pp. 397–415 (2004)
18. Hasanain, M., Elsayed, T., Magdy, W.: Identification of Answer-Seeking Questions in Arabic Microblogs. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM'2014), Shanghai, China, pp. 1839–1842 (2014)
19. Kurdi, H., Alkhaider, S., Alfaifi, N.: Development and Evaluation of a Web Based Question Answering System for Arabic Language. In: Proceedings of the Computer Science & Information Technology CS & IT-CSCP, pp. 187–202 (2014)
20. Khaled, S.: A Survey of Arabic Named Entity Recognition and Classification. Computational Linguistics, 40(2), pp. 469–510 (2014)

21. Kurdi, H., Alkhaider, S., Alfaifi, N.: Development and evaluation of a web based question answering system for Arabic language. International Journal on Natural Language Computing (IJNLC), Volume 3, Issue 2, pp. 11–32 (2014)
22. Ji, H., Grishman, R.: Knowledge Base Population: Successful Approaches and Challenges. In: Proceedings of the49th Annual Meeting of the Association for Computational Linguistics, USA, pp. 1148–1158 (2011)
23. Lopez, V., Uren, V., Sabou, M., Motta, E.: Is Question Answering fit for the Semantic Web?: a Survey. Semantic Web Journal, Volume 2, Issue 2, pp. 125–155 (2011)
24. Lupkowski, P., Leszczyńska-Jasion, D.: Generating cooperative question-responses by means of erotetic search scenarios. Journal of Logic and Logical Philosophy, Volume 24, pp. 61–78 (2014)
25. Shaheen, M., Magdy-Ezzeldin, A.: Arabic Question Answering: Systems, Resources, Tools, and Future Trends. Computer Engineering and Computer Science Arabian Journal for Science and Engineering, Volume 39, Issue 6, pp. 4541–4564 (2014)
26. Prager, J. M., Chu-Carroll, J., Czuba, K., Welty, C., Ittycheiach, A., Mahindru, R.: IBM's PIQUANT in TREC 2003. In: Proceedings of the 12th Text REtreival Conference (TREC 2003), Gathersburg, MD, pp. 283–292 (2003)
27. Pum-Mo, R., Myung-Gil, J., Hyun-Ki, K.: Open domain question answering using Wikipedia-based knowledge model. Information Processing & Management, Volume 50, Issue 5, pp. 683–692 (2014)
28. Surdeanu, M.: Overview of the TAC2013 Knowledge Base Population Evaluation: English Slot Filling and Temporal Slot Filling. In: Proceedings of Text Analysis Conference (TAC) (2013)
29. Trigui, O., Belguith, H. L., Rosso, P.: An automatic definition extraction in Arabic language. In: Proceedings of the 15th International Conference on Applications of Natural Language to Information Systems (NLDB 2010), Cardiff, UK, pp. 240–247 (2010)
30. Trigui, O., Belguith, H., Rosso, P.: DefArabicQA: Arabic Definition Question Answering System. In: Proceedings of the Workshop on Language Resources and Human Language Technologies for Semitic Languages, 7th LREC, Valletta, Malta (2010)
31. Trigui, O.: How to extract Arabic definitions from the Web? Arabic Definition Question Answering System. In: Proceedings of the 16th international conference on natural language processing and information systems (NLDB 2011), Alicante, Spain, pp. 318–323 (2011)
32. Voorhees, E. M.: Overview of the TREC 2003 Question Answering Track. In: Proceedings of the 12th Text REtrieval Conference (TREC 2003), pp. 54–68 (2003)
33. Voorhees, E. M.: Overview of the TREC 2004 Question Answering Track. In: Proceedings of the 13th Text REtreival Conference (TREC 2004), pp. 52–62 (2004)
34. Yang, M. C., Duan, N., Zhou, M., Rim, H. C.: Joint Relational Embeddings for Knowledge-based Question Answering. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, pp. 645–650 (2014)
35. Yahya, A., Salhi, A.: Arabic Text Categorization Based on Arabic Wikipedia. ACM Transactions on Asian Language Information Processing (TALIP), Volume 13, Issue 1 (2014)