Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 96 (2016) 275 – 284
Procedia Computer Science

AQA-WebCorp: Web-based Factual Questions for Arabic

Wided BAKARI[a*], Patrice BELLOT[b] Mahmoud NEJI[c]

[a]*Faculty of Economics and Management, 3018, Sfax Tunisia, MIR@CL, Sfax, Tunisia*
[b]*Aix-Marseille University, F-13397, Marseille Cedex 20, LSIS, Marseille, France*
[c]*Faculty of Economics and Management, 3018, Sfax Tunisia, MIR@CL, Sfax, Tunisia*

Abstract

Working with corpus construction becomes an interesting alternative to different applications of natural language processing, such as, question-answering, machine translation, information retrieval, etc. Similarly, with the heterogeneous data and the user demands for the accurate information, many studies have accentuated the need of the Web to highlight the corpus construction. As well as, Arabic doesn't have an equivalent number of linguistic corpuses as compared to other languages like English. In this paper, we focus on building our corpus of Arab questions-texts. We present a method for recovering text passages. This method is based on a real automatic interrogation of Google, in order to generate passages of texts and answer the factual questions. The first part of this paper describes the formal details about this method; the second part presents some experiments and results that validate our method.

Peer-review under responsibility of KES International

Keywords: Arabic, Corpus, Question analysis, Passage, Google, Corpus construction.

1. Introduction

A corpus is a collection of pieces of texts in electronic forms, selected according to external criteria for end to represent, if possible, a language as a data source for linguistic research [1]. Indeed, a definition that is both specific and generic of a corpus according to [2] is the result of choices that brings the linguists. A corpus is not a simple object; it should not be a mere collection of phrases or a "bag of words". This is in fact a text assembly that can cover many types of text. The construction of a corpus is not an easy task; it is a task that both essential and delicate. Also, it is complex because it depends in large part a significant number of resources to be exploited. One way to diminish this problem is using the Web as data sources. Indeed, the Web is a colossal quantity of texts was recovered freely [5]. It contains billions of text words that can be used for any kind of linguistic research [6]. Additionally, with the internet development and its services, the web has become a great source of documents in different languages and different areas. This source is combined with storage media that allow the rapid construction

* Corresponding author. Tel.: +21696286145;
E-mail address: wided.bakkari@fsegs.rnu.tn

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2016.08.140

of a corpus [3]. In addition, using the Web as a base for the establishment of textual data is a very recent task. The recent years have taken off work attempting to exploit this type of data. From the perspective of automated translation in [4], the others study the possibility of using the websites which offering information in multiple languages to build a bilingual parallel corpus.

Consequently, with the development of electronic media and the heterogeneity of Arabic data on the Web, the idea of building a clean corpus for certain applications of natural language processing, including machine translation, information retrieval, question answer, become more and more pressing. Arabic is also an international language, rivaling English in number of native speakers. However, little attentions have been devoted to this language. Although there have been a number of investigations and efforts invested the Arabic corpus construction, especially in Europe; progress in this area is still limited. In fact, there are few publicly available corpuses, especially for Arabic. The lack and / or the absence of corpus in Arabic have been a problem for the implementation of natural language processing. This also has a special interest in the track of the question answering. Today, the Web has been a driving force in innovations within information retrieval, as users worldwide use search engines to find relevant content on the web. For question-answering, information retrieval methods are used for retrieving documents relevant to the question, and selecting documents likely containing the answer. Most question-answering systems use existing search engines.

In our research, we completed building our corpus of questions-texts AQA-WebCorp (Arabic Question-Answering Web Corpus) by querying the search engine Google. Google has been working on several initiatives to help increase Arabic-language content. Notably, we are concerned our aim, a kind of, giving a question, analyzing texts at the end to answer this question. Therefore, we seek to create and develop our own corpus of pair's questions-texts. This constitution then will provide a better base for our experimentation step. Thus, we try to model this constitution by a method for Arabic insofar as it recovers texts from the web that could prove to be answers to our factual questions. So that, this paper is organized into six sections as follows: it begins with an introduction, followed by the challenges of the Arabic language. Section 3 outlines the earlier work in Arabic; Section 4 shows our proposed method to build a corpus of pairs of questions and texts; Section 5 describes an experimental study of our method; a conclusion and future work will conclude this article.

2. The challenges of the Arabic language

Although, Arabic is within the top ten languages in the internet, it lacks many tools and resources. Meanwhile, Arabic language is the official language in all Arab nations as Tunisia, Egypt, Saudi Arabia and Algeria. Moreover, it is also an official language in non-Arab countries as Chad and Eritrea. The Arabic does not have capital letters compared the most Latin languages. This issue makes so difficult the natural language processing, such as, named entity recognition. Unfortunately there is very little attention given to Arabic corpora, lexicons, and machine-readable dictionaries [20]. In their work [21], the authors suggest that the developed Arabic question-answering systems are still few compared to those developed for English or French, for instance. This is mainly due to two reasons: lack of accessibility to linguistic resources and tools, such as corpora and basic Arabic NLP tools, and the very complex nature of the language itself (for instance, Arabic is inflectional and non concatenative and there is no capitalization as in the case of English). On their part, [22] illustrate some difficulties of Arabic. This language is highly inflectional and derivational, which makes its morphological analysis a complex task. Derivational: where all the Arabic words have a three or four characters root verbs. Inflectional: where each word consists of a root and zero or more affixes (prefix, infix, suffix). Arabic is characterized by diacritical marks (short vowels), the same word with different diacritics can express different meanings. Diacritics are usually omitted which causes ambiguity. Absence of capital letters in Arabic is an obstacle against accurate named entities recognition. And then, in their survey [23], the authors emphasize that as any other language, Arabic natural language processing needs language resources, such as lexicons, corpora, treebanks, and ontologies are essential for syntactic and semantic tasks either to be used with machine learning or for lookup and validation of processed words.

3. Corpus construction from the Web: Case of the Arabic

The corpus is a resource that could be very important and useful in advancing the various language applications. This resource has gained much attention in NLP; it helps the researchers to avoid linguistic generalizations based on his internalized cognitive perception of language. Furthermore, with a corpus, the qualitative and quantitative linguistic research can be done in seconds; this saves the time and the effort. Consequently, the empirical data analysis can help the researchers not only to precede the effective new linguistic research, but also to test the existing theories. Indeed, the task of building a corpus of textual resources from the web is somewhat recent. Anyway, note that few trends emerged in the Arabic corpus construction field and the attempts to exploit this type of data are limited. Although there has been some effort in Europe, which led to the successful production of some Arabic corpus; Progress in this field is still limited. According to [8], the progress has been hampered by the lack of effective corpus analysis tools, such as taggers, stemmers, readable dictionaries to the machine, the corpus viewing tools, etc., that are required for build and enrich a corpus as a research tool. However, several methods have been explored based on the web for other languages, such as the English.

At this point, many researchers have emphasized the importance of a corpus and the need to work on their construction. For example, the work of Resnik studied the possibility of using the websites and offered the information's in multiple languages to build the bilingual parallel corpora [4]. Ghani and his associates performed a study of building a corpus of minority languages from the web by automatically querying the search engines [10]. In order to study the behavior of predicate nouns that highlight the location and movement, the approach proposed by Isaac and colleagues developed a software for the creation of a corpus of sentences to measure if the introduction of prepositions in queries, in information retrieval, can improve the accuracy [7]. Even more, the work of [12] introduced the "BOOTCAT Toolkit". A set of tools that allow an iterative construction corpus by automated querying the Google and terminology extraction. Although it is devoted to the development of specialized corpora, this tool was used by [13] and [14] to the generalized corpus constitution. Similarly, the work of [3] described a tool of building a corpus for the Arabic. This corpus automatically collected a list of sites dedicated to the Arabic language. In another approach of Elghamry, the author proposed an experiment on the acquisition of a corpus from the web of the lexicon hypernymy-hyponymy to partial hierarchical structure for Arabic [15]. Within the framework of the automatic summarization, Maâloul et al. studied the possibility of extracting Arabic texts of the website "Alhayat" by selecting newspaper articles of HTML type with UTF-8 encoding [16]. For his part, [9] showed that the contribution of a corpus in a linguistic research is a huge of many ways. In fact, as such a corpus provides an empirical data that enables to form the objective linguistic declarations rather than subjective. In addition, Ghoul provided a grammatically annotated corpus for Arabic textual data from the Web, called Web Arabic corpus [17]. This corpus consists of 207 356 sentences and 7 653 881 words distributed on four areas: culture, economy, religion and sports. To conclude, in [18], the authors presented arTenTen, an Arabic explored corpus from the web. It is a member of the family of Tenten corpus [19].

Although the text corpus building efforts are focused on English, Arabic corpus can also be acquired from the Web which is considered as a large data source. Most studies in Arabic corpus construction are designed for areas other than the question-answering. However, we also note significant efforts mainly for the question-answering. In this regard, the major of our knowledge, the number of corpus dedicated to Arabic question-answering is somewhat limited. Among the studies that have dedicated to this field, we cite [29] who built a corpus of definition questions dealing the definitions of organizations. They use a series of 50 organization definition questions. They experienced their system using 2000 extracts returned by the Google search engine and Arabic version of Wikipedia.

4. Proposed method for the AQA-WebCorp construction:

The need to have a corpus is a necessity for some processing applications of the Arabic language. As we already mentioned, this paper presents a method for building our corpus of pair's questions and texts from the web. In this section, we discuss the details of this method. Our proposition is simple and organized. It looks for the web addresses corresponding to each question. Indeed, we extracted from the questions a list of features (keywords, focus, and expected answer type). Then, our framework seeks the list of URLs addresses which match those features. Namely, for each given address we propose to recover the webpage that suits him. In this respect, our

corpus construction tool is an interface between the user request and Google. Specifically, it is a way to query the Google database to retrieve a list of documents. Finally, we performed a transformation of each retaining web page from (".html" →".txt"). Finally, we look for if the answer is found in the correspondent text. The text is considered valid to build our body if it contains this answer. Otherwise, we go to the following URL.

These are two ways that we found in literature to retrieve the information from the web for building a corpus: the first one is to group the data located on known sites [4]. Indeed, this way runs a vacuum cleaner web. This ensures the recovery of the pages from a given address. However, the second method investigates a search engine to select addresses from one or more queries (whose the complexity depends on the engine). Thereafter, recover manually or automatically the corresponding pages from these address [7].

We choose to follow the second method. From a list of questions posed in natural language, we ensure the recovery of the list of corresponding URLs. Then, from these URLs, we propose to recover the related web pages. Eventually, we propose to clean these pages so as to produce the lists of texts that will be the foundation that built our corpus. After building our corpus of pair's questions and texts, we do not keep it to that state. In this respect, a stage of analysis and processing will be looked later to achieve our main objective which is the extraction of an adequate and accurate answer to each question. In this respect, querying a search engine can accelerate the recovery of documents online but requires an offline processing of these documents. We think that this is a better solution, but much more complex is the implementation of a Linguistic search engine for a particular purpose. In the rest of this paper, we show in detail our suggestion to build our corpus of pair's questions and texts from the web as well as our experimental results. In our case, the size of the corpus obtained depends mostly on the number of questions asked and the number of documents selected for each question.

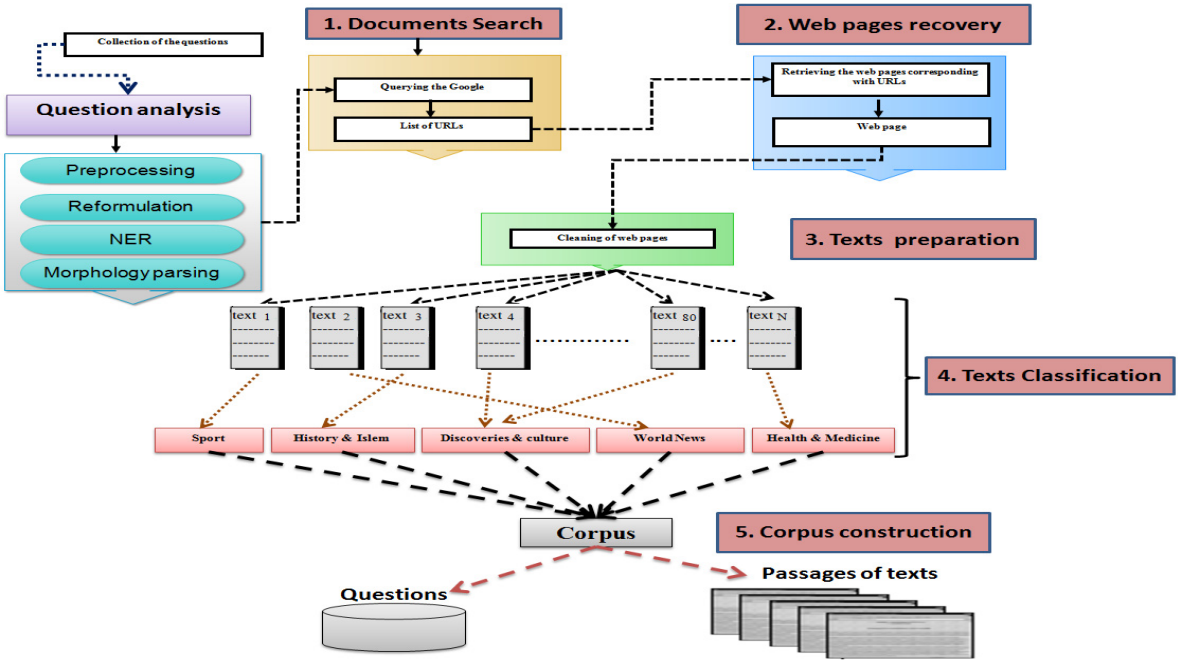

Figure 1proposed method for AQA-WebCorp Construction

4.1. Collecting the questions

We collected the questions that can be asked in different fields, including sport, history & Islam, discoveries & culture, world news, health & medicine. Currently, our corpus consists of 250 pairs of questions and texts. 25 questions translated from TREC, 25 questions translated from CLEF, 100 questions gathered from the forums and 100 from the FAQs. To build our corpus, we used the Arabic texts available on the Internet that is collected being based on the questions posed at the outset. The data collected from the web of the questions and the texts will help us to build an extensible corpus for the Arabic question-answering. Indeed, the collection of these questions is carried out from multiple sources namely, discussion forums, frequently asked questions (FAQ), some questions translated from the two evaluation campaigns TREC and CLEF (Figure 1). The questions are then subjected to an analysis step which assumes a preprocessing and transformation sub step.

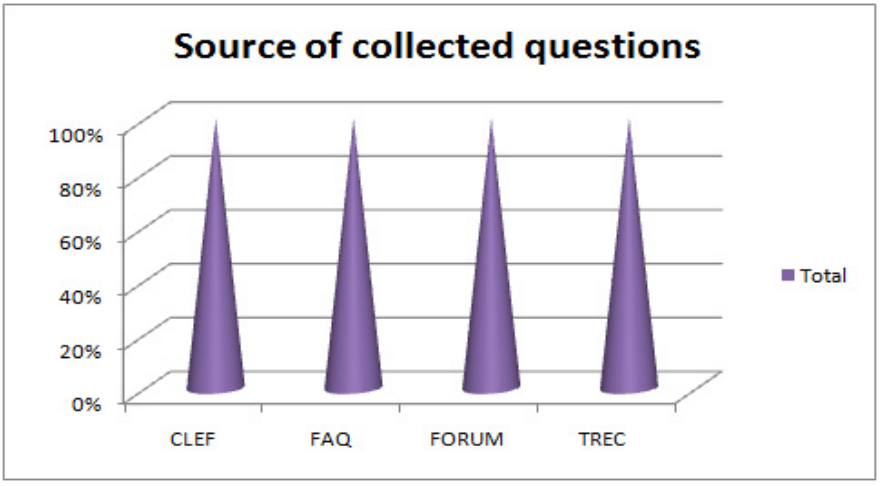

Figure 2 source of the questions used for our corpus

We describe the process of analysis with examples of 5 types of questions collected in our corpus (figure 3). First, for each type of question, we extracted the corresponding characteristics (the focus, the expected answer type and the list of keywords) that are used for the process of document retrieval, analysis and selection of the specific answer. Then we turn the question into their declarative from. It is, then, transformed into a logical representation.

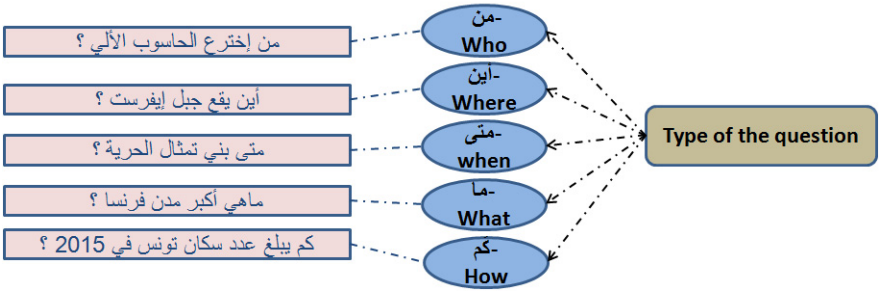

Figure 3 examples of questions used in AQA-WebCorp

4.2. Analyzing the questions

After collecting the questions we have employed an analysis step. For instance, analysis a question is a primordial step in the processing chain of a question-answering system. Several studies emphasize that the task of

extracting an answer to a given question, essentially, requires a deep analysis of the issue [24]. This analysis extracted the key indicators of the question, namely, the expected answer type , the object of the question (focus), the terms that will be used later in the search of the documents that could be prove the answers [25]. These features could be also useful in the next steps when searching the accurate answers [26]. A further essential and complementary purpose for this step is to identify the named entities in the input questions and address the relationships that link those entities.

In addition, our question analysis presents various characteristics that can help us to select the precise answer. This phase is a succession of four steps; the result of each step will be used by the next. It is defined as a preliminary step in the research process of the accurate answers to questions in natural language. In general, the extracted features of this stage, including the keywords, focus, type of response expected facilitate the extraction of specific answer. All information obtained by this module refers to the following stages of the system [26]. In Arabic, the majority of studies focuses on the extraction of keywords and named entity recognition. In our case, we add the reformulation of the question in a declarative form. It is used soon to generate the logical forms.

Consequently, our question analysis module assumes each question to be as a simple declarative sentence that is composed of a sequence of words. It seeks for each sentence the focus as evidence helpful to extract the precise answer. The question analysis is known as a primary task in the processing chain of the question-answering. Therefore, the objective of this step is to obtain the characteristics of the question that could be useful in the following stages of the answer proceedings. In addition, we describe the analysis process with examples of five types of questions collected in our corpus (figure 3). First, for each type of question, we extracted the corresponding characteristics: focus, expected response type and the list of keywords. Then, we transform the question into their declarative form. This latter is transformed into a logical representation.

4.3. Building the AQA-WebCorp

We assume in input a factual question in Arabic language. For each collected question, whose features are extracted, we developed a java script that can extract from certain features the list of html pages corresponding. Then we cleaned these pages to the extent of having a data base of texts that can build our corpus of pair's question-texts. Our method is simple, robust and implemented in Java. Similarly, the object of this method is to induce a tool of generating the passages of texts from the Web in order to answer the questions. In this respect, our corpus construction tool is an interface between the user request and Google. Specifically, it is a way to query the Google database to retrieve a list of documents. The principle of this method is based on four stages, relatively dependent. As illustrated in the (figure 1), the constitution of our corpus of pairs of Arabic question and texts is actually done by developing all of these four steps. We describe in the following each of these steps (Documents research, Web pages recovery, Texts preparation, Texts classification).

4.3.1 Documents research

In our case, to look for an answer to a question in Arabic, we propose to use a search engine (i.e., Google) to retrieve the documents related to each question. Then add post linguistic treatments to those documents which are actually constitute our corpus to have an accurate and appropriate answer. In this respect, querying a search engine accelerates the recovery of documents online but requires an offline processing of these documents. We think that this is a better solution, but much more complex is the implementation of a Linguistic search engine for a particular purpose.

At this stage, the module of documents search is implemented. First, when a question is asked, our tool submits it to the search engine (Google) to identify the list of URLs based on a list of words constitute this question.

4.3.2 Web pages recovery

Consider the following example: our tool can then from the question: « من صمم برج ايفل ؟ » Generate a list of equivalent URLs addresses. The default Web access means is through a search engine such as Google. In addition,

by clicking the "search URLS" button, a list of addresses will be automatically displayed. While for each URL, this prototype can retrieve the necessary information's (host, query, protocol, etc.).

Once the list of URLs is generated, our tool must determine for each address the corresponding web page. This is to look for the corresponding HTML page for each given URL. The following figure illustrates this case. From the address retained in the first step, a set of web pages is recovered. Each of these pages is exported in the format ".html".

4.3.3 Texts preparation

The aim of this step is to transform every web page obtained in the previous step into a ".txt" format. The texts being in ".html" format, and given that the intended application is the statistical language modeling, it seems justified to put them in the ".txt" format. For this, we remove all the HTML tags for each retrieved pages. As we have said before, our method seeks answers to each question in each generated text. It is possible either is to keep the text for own corpus construction work, or to disregard it.

4.3.4 Texts classification

At the moment, the text classification has been done manually according to the topic of the text and the focus for each question. In addition, we are currently extracted a corpus dedicated to the Arabic question-answering. The size of the corpus is in the order of 250 pairs of questions and texts. This is collected using the web as a source of data. The data collected, of these questions and texts from the web, will help us to build an extensible corpus for the Arabic question-answering. The pairs of questions-texts distributed on five areas "إكتشافات; التاريخ والإسلام; أخبار العالم صحة و طب; رياضة; ثقافةو".

4.4. Linguistic analysis of AQA-WebCorp corpus

Data extracted from the Web is not in a suitable textual form to be presented directly to the user. Moreover, to increase the chance to extract the accurate answer, we are implementing step by step the stages of our proposed approach [11]. Then, we add post linguistic processing to those texts which actually constituting our corpuses to have an accurate and appropriate answer. Also, we use a set of steps in order to get the text in the appropriate form; each text of our corpus follows those steps. One of the advantages of corpora is that they can readily provide quantitative data which intuitions cannot provide reliably. The following stage of linguistic analysis of our corpus is to analyze the 250 texts generated from each question. This stage accepts as input an Arabic text in txt format and generates an annotated and analyzed text. The analysis consists of several sequential sub steps. First, preprocessing is performed to eliminate the stop words. Then, a segmentation step determines the division of text into tokens (sentences). Indeed, the study of [31] shows that a text analysis without segmentation lead to unreliable results. Therefore, segmentation plays a very important role in most applications of automatic natural language processing (ex. information extraction, automatic summarization, etc.). Afterward, the recognition of named entities is performed by ArNER [28] to determine the set of named entities. Finally, In order to identify the grammatical category of the words belong in the text; we use Alkhalil parser [30] to carry out the morphological analysis. This stage is an essential step in achieving most applications in automatic language processing.

5. Evaluation

A principal purpose for creating the corpus was to aid in the development of an Arabic question-answering system. In this section, we describe our experiments results to obtain a reliable Arabic corpus of questions and texts. We test our proposed method with a set of collected questions which consists of 250 factual questions. For each URL, we automatically retrieved the corresponding Web-page using a script java and we have removed HTML tags and special characters as well as spelling mistakes were also corrected.

As an example, consider the following question « من صمم برج ايفل ؟», the proposed tool for generating the passages of the texts from the Web attributes the URL corresponding addresses to it. While for each URL, this prototype can retrieve the necessary information's (host, query, protocol, etc.). While for each URL, this prototype can retrieve the necessary information's (host, query, protocol, etc.). Next, it determines for this address the corresponding web page. This is to look for the corresponding HTML page for each given URL. Finally, this prototype assumes the transformation of every obtained web page into a ".txt" format. For this, we have removed the HTML tags and the special characters. Figure 4 shows a passage of text corresponding to this question.

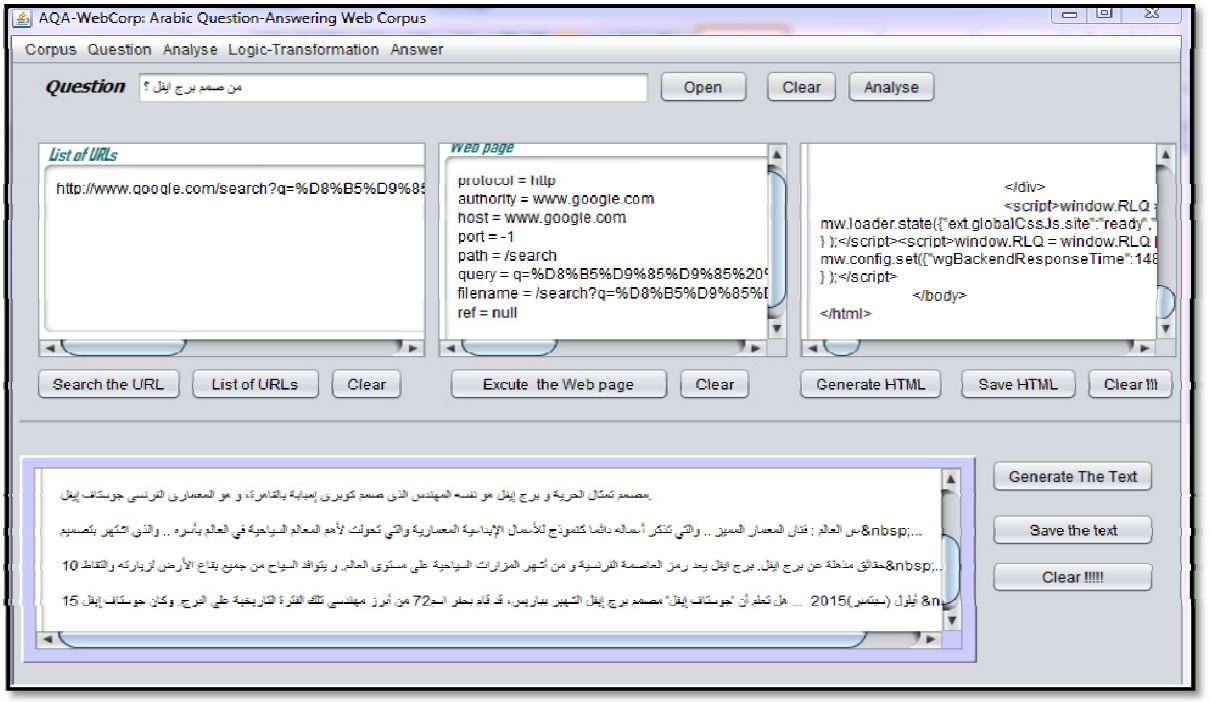

Figure 4 text extracted from the Web for the following question: " من صمم برج ايفل ؟

For the experiments, we have accomplished two experiments over the corpus. The first one is with 115 questions and the next is with 250 questions. We have used the Accuracy and c@1 measures. These measures are used to measure the passages of texts automatically generated from the web that could answer those questions. Our final corpus is now consisted of 250 pairs of questions-texts. However, it should be noted that there are some questions that are either not answered, either incorrectly answered. In [27], the authors mentioned that not answering has more value than answering incorrectly. So, to evaluate those unanswered questions, we use the c@1 measure. It is an extension of accuracy measure (the proportion of correctly answered questions) with this feature and a very easy to understand interpretation. This measure has a good balance of discrimination power, stability and sensitivity properties. Two experiments were also carried out on the proposed corpus to evaluate the method presented in this paper. The results obtained were very promising. We demonstrated the completeness and the representatives of our corpus using those metrics and show its suitability to answer the questions.

CA: Number of Correctly Answers.
UQ: Number of Unanswered Questions.
TQ: Total of Questions.

The Accuracy measures the number of questions correctly answered divided by the total number of collected questions (correctly answered and not correctly answered).

$$Accuracy = CA / TQ \quad (1)$$

The c@1 measures the proportion of correctly answered questions.

	preliminary experiment	final experiment
Total of questions	115	250
Correctly answered	101	224
Unanswered = (incorrectly+not(answered))	14	26
Accuracy	0.87	0.89
c@1	0.98	0.89

$$C@1 = (CA + UQ * (CA / TQ)) / TQ \quad (2)$$

Table 1. Results of the two experiments of AQA-WebCorp

6. Conclusion and perspectives

Nowadays, it is difficult to find a corpus designed for the processing of the natural language processing and, more specifically, for the Arabic language. In this paper, we have created a new Arabic corpus for the question-answering. The data for the proposed corpus were collected from several sources. This data is consisted of 250 pairs of questions-texts. We have chosen to use the Web as a source of the texts because it is essentially an enormous database of mostly textual documents and it offers great opportunities to corpus construction.

For further work, we will continue in this line of research by improving the logic representation stage. Moreover, the logic representation was driven mainly by the transformation of textual data to the predicate-argument form.

Acknowledgements

I give my sincere thanks for my collaborators Professor Patrice BELLOT (University of Aix Marseille, France) and Mr. Mahmoud NEJI (University of Sfax-Tunisia) that I have benefited greatly by working with them.

References

[1] Sinclair, J. (2005). Corpus and text - basic principles. In M. Wynne (Ed.), Developing linguistic corpora: A guide to good practice (pp. 1–16). Oxford, UK: Oxbow Books.

[2] Rastier F. (2005). « Enjeux épistémologiques de la linguistique de corpus », in : Williams C. G. (dir), La linguistique de corpus, Rennes : P.U.R

[3] Meftouh K., Smaïli K. and Laskri M.T., (2007). Constitution d'un corpus de la langue arabe à partir du Web. CITALA '07. Colloque international du traitement automatique de la langue arabe. Iera, Rabat, Morocco, 17-18 juin 2007.

[4] Resnik, P., (1998), Parallel strands: A preliminary investigation into mining the web for bilingual text, in conference of the association for machine translation in the Americas, 1998.

[5] GATTO, M., (2011), The 'body'and the 'web': The web as corpus ten years on.ICAME JOURNAL, 2011, vol. 35, p. 35-58.

[6] Kilgarriff, A., & Grefenstette, G. (2001, March). Web as corpus. In Proceedings of Corpus Linguistics 2001 (pp. 342-344). Corpus Linguistics. Readings in a Widening Discipline.

[7] Issac, F., Hamon, T., Bouchard, L., Emirkanian, L., and Fouqueré, C., (2001), extraction informatique de données sur le web : une expérience, in Multimédia, Internet et francophonie : à la recherche d'un dialogue, Vancouver, Canada, mars 2001.

[8] Atwell, E., Al-Sulaiti, L., Al-Osaimi, S. & Abu Shawar, B. (2004). A review of Arabic corpus analysis tools. In B. Bel & I. Marlien (Eds.), Proceedings of TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles (volume 2, pp. 229–234). ATALA.

[9] Mansour, M. A., (2013), The Absence of Arabic Corpus Linguistics: A Call for Creating an Arabic National Corpus, International Journal of Humanities and Social Science, Vol. 3 No. 12 [Special Issue – June 2013].

[10] Ghani, R., Jones, R., Mladenic, D., (2001), Mining the web to create minority language corpora, CIKM 2001, 279-286.

[11] Bakari, W., Trigui, O., and Neji, M. (2014, December). Logic-based approach for improving Arabic question answering. In Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on (pp. 1-6). IEEE.

[12] Baroni, M., Bernardini, S., (2004), Bootcat: Bootstrapping corpora and terms from the web, proceeding of LREC 2004, 1313-1316.

[13] Ueyama, M., Baroni, M., (2005), Automated construction and evaluation of Japanese web-based reference corpora, proceedings of corpus linguistics, 2005.

[14] Sharoff, S., (2006), Creating general–purpose corpora using automated search engine, in Wacky! Working papers on the web as corpus, Bologna: GEDIT 2006, 63-98.

[15] Elghamry, K. (2008). Using the web in building a corpus-based hypernymy-hyponymy Lexicon with hierarchical structure for Arabic. Faculty of Computers and Information, 157-165.

[16] Maâloul, M.H., Keskes, I., Belguith, L.H., and Blache, P., (2010), "Automatic summarization of arabic texts based on RST technique",12th International Conference on Enterprise Information Systems (ICEIS'2010), 8 au 12 juin 2010, Portugal.

[17] Ghoul, D., (2014), Web Arabic corpus: Construction d'un large corpus arabe annoté morpho-syntaxiquement à partir du Web. In ACTES DU COLLOQUE , 2014 (p. 12).

[18] Arts, T., Belinkov, Y., Habash, N., Kilgarriff, A., & Suchomel, V. (2014). arTenTen: Arabic Corpus and Word Sketches. Journal of King Saud University-Computer and Information Sciences, 26(4), 357-371.

[19] Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., and Suchomel, V., (2013), The TenTen Corpus Family. International Conference on Corpus Linguistics, Lancaster.

[20] Hammo B., Ableil S., Lytinen S. and Evens M. (2004). "Experimenting with a Question Answering system for the Arabic language", In Computers and the Humanities. Vol. 38, N°4. Pages 397—415.

[21] Bekhti S., Rehman A., AL-Harbi M. and Saba T. (2011)."AQUASYS: an arabic question-answering system based on extensive question analysis and answer relevance scoring", In *International Journal of Academic Research*; Jul2011, Vol. 3 Issue 4, p45.

[22] Abdelnasser, H., Mohamed, R., Ragab, M., Mohamed, A., Farouk, B., El-Makky, N., & Torki, M. (2014). Al-Bayan: An Arabic Question Answering System for the Holy Quran. ANLP 2014, 57.

[23] Ezzeldin A. M. and Shaheen M. (2012). "A survey of Arabic question answering: challenges, tasks, approaches, tools, and future trends", the 13th International Arab Conference on Information Technology ACIT'2012. Dec.10-13. ISSN 1812-0857.

[24] Embarek, M., (2008), "Un système de question-réponse dans le domaine médical Le système Esculape", PhD thesis. Université de Paris-Est. Juillet 2008.

[25] Zweigenbaum, P., Grau, B., Ligozat, A. L., Robba, I., Rosset, S., Tannier, X., and Bellot, P. (2008). Apports de la linguistique dans les systèmes de recherche d'informations précises.

[26] Rodrigo, Á., Perez-Iglesias, J., Peñas, A., Garrido, G., and Araujo, L. (2010). A Question Answering System based on Information Retrieval and Validation. InCLEF (Notebook Papers/LABs/Workshops).

[27] [Peñas et al., 11] Peñas, A. and Rodrigo, A. A Simple Measure to Assess Non-response. In Proceedingsof 49th Annual Meeting of the Association for Computational Linguistics-Human Language Technologies (ACL-HLT 2011), Portland, Oregon, USA, June 19-24, 2011.

[28] Zribi I., Hammami S. M. and Belguith L. H. (2010) "L'apport d'une approche hybride pour la reconnaissance des entités nommées en langue arabe", *In TALN'2010*, Montréal, 19-23 juillet 2010 (pp. 19–23).

[29] Trigui, O., Belguith, L.H. and Rosso, P., (2010), "DefArabicQA: Arabic Definition Question Answering System", In Workshop on Language Resources and Human Language Technologies for Semitic Languages, 7th LREC. Valletta, Malta. 2010.

[30] Ould Bebah M. O. A, Mazroui. A, Meziane A., Lakhouaja A. (2011). Alkhali Morpho Sys. In International Computing Conference in Arabic. Riadh, Arabie Saoudite.

[31] Ghassan M. (2001) Analyse informatique des signes typographiques pour la segmentation de textes et l'extraction automatique des citations. Réalisation des Applications informatiques : SegATex et CitaRE, PhD thesis, Paris-Sorbonne.