High-Performance Arabic Question Answering System

HAMADENE Abdelbaki
Alexandria, Egypt

Résumé
Question Answering systems are emerged to provide unique, brief and concise answers instead of links toward web pages returned by search engines. Many systems have been deeply thought and efficiently developed for other natural languages especially English. Research studies to develop efficient Arabic Question Answering Systems are still few and need important care to many aspects and Arabic particularities (short vowels, absence of capital letters, complex morphology, etc.). The proposed Question Answering system (ArQA) with sophisticated architecture deals with Arabic factoid questions. It employs the latest approaches of NLP, IR and question answering systems. The experimental results show the effectiveness of our approach based on N-grams matching, question's focus filtering, and studies about text similarity scoring. Using standard metrics (adopted by TREC), ArQA achieves an overall accuracy of more than 86% and outperforms other systems.

1. Introduction

Since the time of the earliest digital computers, the amount of information available in the form of electronic text has grown exponentially, especially due to computer networks. Although search engines are an effective solution for finding documents matching a user query, they are less efficient for finding a precise data. It is the ambition of the Question Answering (QA) systems to find this data. QA is a challenging task. It involves state of the art techniques in various fields such as Information Retrieval (IR), Natural Language Processing (NLP), Artificial Intelligence (AI), Managing large data sets, and Advanced Software Technologies.

This paper presents ArQA, a new QA System for the Arabic language. It shows how the use of a sophisticated architecture influences the effectiveness of Arabic QA. Several NLP and IR tools are used to get a more accurate system. The normalization phase was simplified to keep data consistency. The determination of answer type was based on new set of Arabic rules. Keywords were expanded using stems and Arabic dictionary of synonyms (except for proper nouns). The stemming phase was conditioned by disabling Named Entities processing. A simple keyword-based IR system was implemented without ranking step to remove its overhead. Answers were extracted based on N-grams similarity function (N=2), and on question's focus-based filtering function. The validation module used special set of rules and formulas in the scoring stage. The remainder of this paper is organized as follows. Section 2 gives a brief background of QA and the Arabic language. Section 3 describes the state of art and some existing approaches for QA systems for both Arabic and other languages. Section 4 explains the different modules of the proposed system. Section 5 contains the experimental results with its analysis in section 6, while conclusion and future work are the subject of section 7.

2. Background

The Question Answering (QA) systems can be defined as sophisticated IR systems that can return a direct, short and precise answer, or a passage containing the answer to a user request instead of a search engine that returns a set of documents deemed relevant. The QA systems have a complex architecture and rely on many search techniques. The question is processed to extract answer without manual intervention.

2.1 Question Answer System Architecture

Although the techniques differ from one system to another, most QA systems are based on architecture of four modules (Figure

1). These modules are mainly based on techniques of NLP and IR. The IR tools are used mainly in the search of the most relevant documents and passages, while NLP techniques can improve the IR procedures by offering the possibility of a deeper analysis of the question and documents.

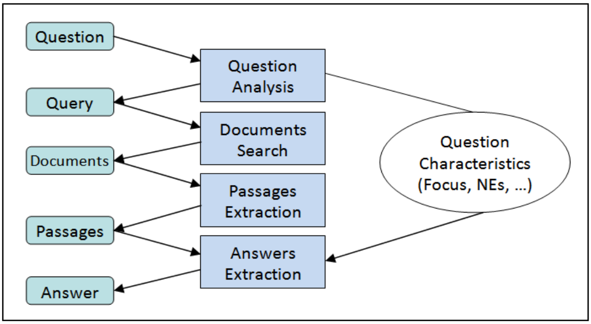

Figure 1. Generic Architecture of Question Answer System

- The first module concerns the analysis of the question. It aims specifically to extract the maximum of information from the question, such as the type of question, the question's focus, the type of expected answer and the question keywords.
- The second module aims to mine a corpus of texts to select a set of documents or extracts of documents and facilitating the processing of the following chain.
- The third module is responsible for analyzing the selected documents and extracting the candidate passages that may contain the answer.
- Finally, the fourth module allows us to search the answer in the selected passages, depending on the question.

2.2. Arabic Language and its Challenges

Arabic is considered one of the six official languages of the United Nations and the mother language of more than three hundred million people. It has a special status as the formal written standard of the media, culture and education across the Arab World. It is the religious language of all Muslims of various ethnicities around the world. Arabic, the language, is written from right to left using the Arabic script. It is a Semitic language with 28 alphabet letters (25 consonants and 3 long vowels).

The need for information in Arabic is quite high. This language has recently become the focus of an increasing number of projects in NLP and Computational Linguistics. Many aspects slow down progress in Arabic NLP compared to the accomplishments in English and other European languages [2]. These aspects include:

1. The Arabic language is highly derivational and inflectional

- Arabic is a derivational language: derivations in Arabic are almost always moulded, thus: Lemma = Root + Pattern. To find an Arabic word in a dictionary, first its root is extracted and then this root is searched. This is because the vocabulary of Arabic is essentially built from roots derivation by adding affixes (prefix, infix, or suffix) to the root according to several patterns that are around 120 [3]. Finding the root of a word is to identify its morphemes (Stem): the affixes and the root. A stem can be composed of one part [root], as: (ع م ل); two parts [root + pattern], as : (ع ا م ل) or three parts [root + pattern + affixes], like: (فا ل ع ا م ل و ن). Arabic has about ten thousand roots. Figure 2 shows an example of Arabic derivation.

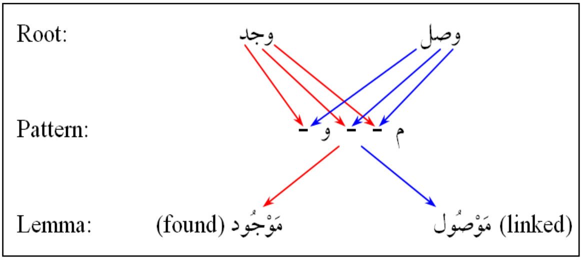

Figure 2. An example of Arabic language derivation

- Arabic is an inflectional language: word = root + affixes (prefix, infix, and suffix). Both prefixes and suffixes are allowed to be combinations, and thus a word can have zero or more affixes. (Figure 3). Generally, relevant documents are those containing the query keywords. However, if keywords appear in a document with additional inflections this document would be classified as irrelevant even if it contains the correct information.

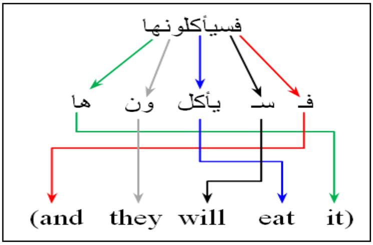

Figure 3. An example of Arabic words composition

1. The absence of diacritics creates ambiguity and therefore, complex morphological rules are required to identify the tokens and parse the text.
2.Capital letters are a crucial characteristic to be used in the recognition of Named Entities. Capitalization is not used in Arabic, which makes it hard to identify proper names, acronyms, and abbreviations.

3. Arabic writing direction is from right-to-left. Moreover, numbers are written from left to right which is a real challenge for the Arabic text editors to handle words written from left to right and others from right to left in the same line.
4. In addition to the above issues, there is a lack of Arabic corpora, lexicons, and e-dictionaries, which are essential to advance research in different areas [10].

3. Related Works

Several developed QA systems for Arabic and other languages exist. The QA systems can be differentiated according to the adopted search strategies. Following is a presentation of the most successful approaches in the task of QA in recent evaluation campaigns TREC (http://trec.nist.gov), CLEF (www.clef-campaign.org) and EQueR [4].

3.1. Non Arabic QA

1. QALC [7] was the first QA system developed for English in the TREC evaluation campaign in 1999. It relies on a set of modules for language processing: questions analysis, documents selection, named entities recognition and answers extraction.
2. QRISTAL [12] is a multilingual QA system developed to retrieve answers from a local database or from the Web. It consists of several modules of NLP, including parsing, and semantic disambiguation. Due to massive use of NLP tools, QRISTAL was ranked first over seven participating systems in EQueR 2004.
3. PIQUANT [6] is based on the use of multiple QA systems as the type of question and therefore has greater relevance through the plurality and redundancy of found answers. Thus, PIQUANT is based on statistical and several NLP tools.
4. AnswerBus (http://www.answerbus.com/) is an open-domain QA system based on sentence level Web IR. Five search engines and directories are used to retrieve relevant Web pages

to user questions. Its rate of correct answers to TREC-8’s 200 questions was 70.5% with the average response time being seven seconds.

3.2. Arabic QA

Research in Arabic QA systems is not new but its results are not accurate yet. Few research works (systems) have been developed for Arabic QA such as:

1. AQAS [13] is a knowledge-based QA system. It extracts answers only from structured data and not from raw text. There have been no experiments or results.
2. QARAB [10] uses both IR and NLP techniques. The IR system is constructed using a relational DBMS based on Salton’s vector space model. The NLP system is composed of a set of tools to tag Arabic text, and to identify proper no. It was assumed that answer exists in one document. QARAB's developers claimed that they obtained 97.3% in each of the precision and the recall. These (possibly biased) results are not reliable because such accuracy was not achieved in any other language in the QA state-of-the-art. Moreover, كيف and لماذا (How and Why) questions are not supported.
3. ArabiQA [5] is based on a passage retrieval system (JIRS) module and a NER system. It embeds an Answer Extraction module dedicated especially to factoid questions. Authors developed an Arabic NER system called ANERsys. A Maximum Entropy (ME) approach was employed. A precision amounting to 83.3% was reached. The system was not completed.
4. QAS [8] is developed with the spiral model approach. It differs from most QA systems in its dependency on data redundancy rather than complicated linguistic analyses. The system cannot handle 'how' or 'why' questions. It uses Tagger to analyze questions and documents. The used IR system is implemented as done in QARAB system. QAS get interpolated precisions between 43% and 100%.

5. QASAL [14] deals factoid and definition questions. It employs the NooJ platform. QASAL has an architecture consisting of three modules: Question Analysis, Passage Retrieval, and Answer Extraction module. No results were provided for factoid questions.

Most of these researches cited above, have not made test-bed publicly available, which makes it impossible to compare their evaluation results. They claim to provide acceptable results. Unfortunately, they are not open source or it is very hard to integrate new add-ons to them. Moreover, they failed to get good results as for other languages. These reasons led to investigate this interesting subject to get a system with high performance (in terms of accuracy and response time).

4. Proposed ArQA system architecture

The proposed system [9] is a sophisticated one that has a pipeline architecture consisting of four components: Question Processing, Passages Retrieval, Answers Extraction, and Answers Validation modules. The input of the system is an Arabic question which is processed to output an answer following the sequence diagram in Figure 4.

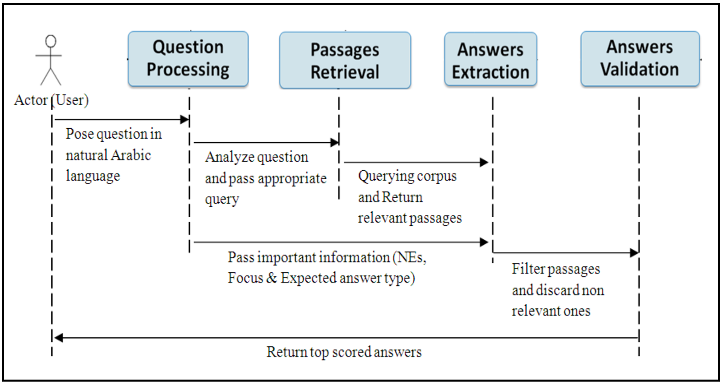
Actor (User)
Question Processing
Passages Retrieval
Answers Extraction
Answers Validation
Pose question in natural Arabic language
Analyze question and pass appropriate query
Querying corpus and Return relevant passages
Pass important information (NEs, Focus & Expected answer type)
Filter passages and discard non relevant ones
Return top scored answers

Figure 4. ArQA sequence diagram

4.1. Questions Processing Module

The question processing is an important step in the QA system. Indeed, it is essential for a system to analyze a question as thoroughly as possible to determine the search strategy to apply. Question analysis is done through following steps

1. Tokenization a tokenizer was developed for splitting words from punctuation.
2. Normalization performed to make the data sets more consistent and to overcome differences between documents and errors in texts. ArQA normalizes as follows

- Replacing initial أ, آ, إ by ا and replacing the sequence ىء by ئ
- Replacing final ى by ي and replace final ة by ه

3. Answer typing provides information on the expected answer type. The type is formalized as a type of named entity (NE) (person, date, place...). ArQA uses simple rules for the beginning of the question: من [Who] for Person, أين [Where] for Location, كم [How many] for Quantity, ما ، ما اسم [What] for Object, متى [When] for Date or Time and لماذا [Why] for Reason.

In Arabic, these interrogative nouns can be expressed in other forms with the same role. Table 1 shows a list of types and forms of questions used in ArQA.

Question type	Expected answer type
متى When	زمان Date or Time
في أي + عبارة زمان (عصر / سنة / تاريخ / ...) In which + Time terms (year/day/ date..)	
أين Where	مكان Location
في أي + عبارة مكان (مكان / بلد / مدينة / ...) In which + Location terms (place/city/ country..)	

Table 1. Example of defined Question Types and Forms in ArQA

4.Named Entity Recognition (NER)

To extract different types of Named Entities (NEs) in the passage. The most common are NEs of type MUC (Message Understanding Conferences).

5.Focus determination

To extract the main entities concerned by the question. Question's focus will be used later in the filtering phase.

6.Keywords Extraction

After elimination of stop-words, the important words in the question (keywords) are extracted to compose a query. The stop lists generally used in Arabic text processing are the Khoja list, which has only 168 words and the 1,131 words translated from English. A new list of more than 10300 words was created to decrease the size of useless proceeded words in the Arabic texts.

7.Keywords Expansion

Exploiting question's keywords does not necessarily allow finding the answer in a document because the meaning of a word can be represented or interpreted in different ways. That's why ArQA extends the query by adding terms semantically related to its keywords by use of an Arabic dictionary of synonyms as lexical resource and semantic knowledge base (instead of heavy Arabic WordNet). NEs are not expanded to avoid ambiguity.

8. Stemming To reduce variant words to single stem to overcome grammatical variations of words. Words in the question and in the sentence are returned to their root form. Therefore, although their conjugations are different, system considers that two words are similar based on the similarity of their stems. Khoja's stemmer [11] was adapted to perform this task. Named Entities are not stemmed.

9. Query generation

   This step consists of the formulation of expanded keywords in Boolean formula. Resulted query will be passed to the next module to retrieve related passages.

4.2. Passages Retrieval Module

This step is crucial because the QA system cannot find an answer if it is not present in the selected documents. Advanced techniques for NLP are too heavy to be used on a large amount of text, so ArQA system uses a simple keyword based IR system to restrict the volume of text to analyze. The corpus is pre-indexed offline to reduce response time. Retrieved passages are not ranked to avoid ranking process overhead.

4.3. Answers Extraction Module

In ArQA, answer is the result of a matching made between the representation of the question and the selected portions of text after analysis of candidate passages. Two methods will be used in the answer extraction

- Filtering computes semantic similarity between the question's focus and the candidate answer to exclude irrelevant passages. However, and differently from other systems, ArQA discarded other sentences only in the case of questions about quantity or date because for questions about persons, places or objects, it's difficult to recognize all existing NEs, which can eliminate relevant sentences wrongly.
- Matching gives a score to the amount of overlap between the content of the question and a potential answer. ArQA uses N-grams similarity function to decide whether tow words are similar or not.

4.4. Answers Validation Module

Validation estimates for each of the candidate answers the probability of correctness and ranks. For ArQA system, answer validation is done through tow tasks

1. Scoring accuracy scoring is calculated by tow formulas. First formula is based on a study of text similarity measuring techniques [1] and the second is based on the question category (Table 2).

General_Score = Score_1 + Score_2 + Score_3 + Score_4
Total_Score = General_Score + Specific_Score_i

Where the four considered basic factors (Score_i) are defined below:

- Score_1 = (NMKS / TNKS) x 100

Whith: NMKS is the number of matching keyword's stems in the candidate sentence and TNKS is the total number of keyword's stems.

- Score_2 = (NMOK / TNOK) x 100

Whith: NMOK is the number of matching keyword in their original forms in the candidate sentence and TNOK is the total number of keyword in their original forms.

- Score_3 = (OKSF / TNKS-1) x 100

Whith: OKSF is the order of keyword's stems factor in the candidate sentence.
The order of keywords factor is a variable incremented by the distance between the two words each time two keywords in a candidate sentence appear in the same order as they do in the user’s question (different scores for active and passive form).

- Score_4 = (SKSF / TNKS-1) x 100

Whith**:** SKSF is the sequence of keyword's stems factor in the candidate sentence. The sequence of keywords factor is a variable incremented by 1 each time two keywords in candidate sentence appear in the same direct succession as they do in the user’s question. The main purpose is to assess the relative importance of each clue.

2. Ranking in ArQA system the ranking process consists in ordering sentences based on their scores (Total_Score) from the highest to the lowest one.

Question type	**Scoring Rules**
متى When	Specific_Score_1 = 10 x (F) F = 2 if sentence contains the main verb in the question F = 0 otherwise If Sentence contains (question Focus), Specific_Score_1 + = 3 If Sentence contains (Time-Expression), Specific_Score_1 + = 4
في أي + عبارة زمان (عصر / سنة / ...) In which + Time terms (year/...)	

Table 2. Example of the question-type related scoring rules used in ArQA

5. Experiments and results

The system was developed using Java programming language. To evaluate ArQA system, the annotated corpus ANERcorp (available online) was adapted (pre-processed) offline (cleaned and indexed). ANERcorp consists of 316 articles from different newspapers and sources to obtain a corpus as generalized as possible (Table 3). ANERcorp contains 150,286 tokens.
Random samples of 20 volunteers of different educational and social levels were asked to pose questions about the corpus and provide the right expected answers. 240 questions of different types (persons, objects, locations, dates, quantities and reasons) were asked to ArQA.

Source	**Ratio**
http://www.aljazeera.net	34.8%
Other newspapers and magazines	17.8%
http://www.raya.com	15.5%
http://ar.wikipedia.org	6.6%
http://www.ahram.eg.org	5.4%
http://www.alalam.ma	5.4%
http://www.alittihad.ae	3.5%
http://www.bbc.co.uk/arabic/	3.5%
http://arabic.cnn.com	2.8%
http://www.addustour.com	2.8%
http://kassioun.org	1.9%

Table 3: Ratio of sources for the extracted articles of ANERcorp

Unlike other systems, in the experiments the numbers of questions were distributed equally between the precedent types to avoid biasing results (Table 4). After posting questions the returned answers were compared to those expected by the users and judged manually. For response time, several runs were done then the average was got.

Question type	Number of questions	Correct answers	Accuracy	MRR	Average Response Time
Person	40	39	97.50 %	0.99	944 ms
Time	40	37	92.50 %	0.93	2066 ms
Objects	40	37	92.50 %	0.93	1341 ms
Quantity	40	34	85.00 %	0.91	1413 ms
Location	40	33	82.50 %	0.84	6078 ms
Reason	40	27	67.50 %	0.73	1727 ms
Overall	**240**	**207**	**86.25 %**	**0.87**	**2262 ms**

Table 3. Overall results obtained with ArQA

Systems	Number of Questions	Corpus	Accuracy	MRR	Average Response Times
ArQA	240	ANERcorp	86.25 %	0.87	2.3 s
AnswerBus	200 (TREC)	Web pages	70.50 %	-	7 s
QArabPro (Arabic)	335	Wikipedia (75articles)	84.18 %	-	-

Table 4. Comparison between ArQA and other QA systems' performances

6. Discussion

Obtained results show a good performance of ArQA system in terms of both accuracy (ratio of number of correctly answered questions / number of questions) and response time which is critical to the usefulness of a QA system. The system achieved 86.25 % overall accuracy with an average response time less than three seconds (2262 ms) on a machine with low specs (CPU: Intel® 1.60 GHz, RAM: 512 MB). Results are better than those obtained in similar systems (Table 5). Best results were for questions about persons because Person Entities represent the greatest percent of the corpus and of the gazetteers. The short response time is due to the simple used rules. For location's questions specific and complex rules were set to look for location terms and prepositions indicating location that's why the response time for this type of questions was the longest one. The low results for (Why) “لماذا” were expected because of the difficulty in handling such questions. This type of question is usually related with causal justification and requires deep semantic processing.

7. Conclusions and future works

In this paper a new Arabic QA system called ArQA was described. This system provides answers to factoid questions expressed in Arabic language. ArQA is a sophisticated system based on the use of the conditioned stemming, the synonymy expansion, the question’s focus filtering, the N-grams matching, and rule-based scoring. In this system, with pipeline architecture, each module is the result of combination of several approaches, methods and tools to improve the accuracy of returned answers. On used test set the system achieved 86.25% overall accuracy within average response time less than three seconds. It is planned to bring ArQA to the internet and to use statistical approach for question classification as well as for the retrieval tasks.

References

[1]Achananuparp, P. Hu, X. Zhou, X. Zhang, X., (2008) “*Utilizing Sentence Similarity and Question Type Similarity to Response to Similar Questions in Knowledge-Sharing Community*”. In: Proceedings of QAWeb 2008 Workshop, Beijing, China.

[2]Al-Daimi, K., and Abdel-Amir, M., (1994) “*The Syntactic Analysis of Arabic by Machine*”, Computers and Humanities, Vol. 28, No. 1, pp. 29-37.

[3]Al-Kharashi, I., Evens, M., (1994) “*Comparing Words, Stems, and Roots as Index Terms in an Arabic Information Retrieval System*”, Journal of the American Society for Information Science and Technology, Vol. 45, No. 8, pp. 548–560.

[4]Ayache, C. Grau, B. and Vilnat, A., (2006) " *EQueR: the French evaluation campaign of question answering system EQueR/Evalda*". In proceedings of the 5th international Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy, pp. 1157-1160.

[5]Benajiba, Y., Rosso, P., Lyhyaoui, A., (2007) "*Implementation of the ArabiQA Question Answering System's components*", Workshop on Arabic Natural Language Processing, 2nd Information Communication Technologies Int. Symposium, ICTIS-2007, Fez, Morocco.

[6]Chu-Carroll, J., Prager, J., Welty, C., Czuba, K., Ferrucci, D., (2002) "*A Multi-Strategy and Multi-Source approach to question answering*", 11th Text Retrieval Conference (TREC-11).

[7]Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G., Jacquemin, C., Masson, N., Lecuyer, P., (2000) "*QALC-the question answering system of LIMSI-CNRS*", Technical report: LIMSI-CNRS TREC 9 evaluation.

[8]Ghassan, K., Awni, H., Riyad, A., Majdi, S., (2009) "*A New Question Answering System for the Arabic Language*", American Journal of Applied Sciences, Vol. 6, No. 4, pp. 797-805.

[9]Hamadene, A., Mohamed, S., Osama, B. (2011) "*ArQA: an intelligent question answering system*". In: Proceedings of Arabic Language Technology International Conference (ALTIC 2011), Alexandria, Egypt, pp. 129-136.

[10] Hammo, B., Abu-Salem, H., Lytinen, S., (2002) "*QARAB: A Question Answering System to Support the Arabic Language*", Workshop on Computational approaches to Semitic languages, ACL, Philadelphia, pp. 55-65.

[11] Khoja, S. Garside, R., (1999) "*Stemming Arabic Text*", Lancaster, UK, Computing Department, Lancaster University.

[12] Laurent, D., Séguéla, P., (2005) "*Qristal, système de Question-Réponse*" *(in French)*, Traitement Automatique des Langues Naturelles (TALN 2005), Dourdan, France.

[13] Mohammed, F.A., Nasser, K., Harb, H.M., (1993) "*A knowledge-based Arabic Question Answering System (AQAS)*", ACM SIGART Bulletin, pp. 21-33.

[14] Wissal, B., Ellouze, M., Mesfar, S., Hadrich Belguith, L., (2009) "*An Arabic Question-Answering system for factoid questions*", IEEE International Conference on NLP and Knowledge Engineering. pp 797-805.