Towards an Arabic Question Answering system

Article · January 2006

CITATIONS
21

READS
1,198

3 authors, including:

Yassine Benajiba
Thomson Reuters
29 PUBLICATIONS **944** CITATIONS

SEE PROFILE

Abdelouahid Lyhyaoui
Ecole de Sciences Appliqués de Tanger
89 PUBLICATIONS **381** CITATIONS

SEE PROFILE

Towards an Arabic Question Answering System

Paolo Rosso[1], Yassine Benajiba[1], and Abdelouahid Lyhyaoui[2]
[1]Natural Language Engineering Lab., RFIA Group
Dept. of SIC, Polytechnic University of Valencia, Spain
[2] Engineering Systems Lab. (LIS), Abdelmalek Essaadi University, Tangier, Morocco
Email: {prosso, ybenajiba}@dsic.upv.es, lyhyaoui@gmail.com

Abstract – **The Question Answering (QA) task is a field of Natural Language Processing (NLP) in which researchers try to find ways to make investigation easier for web users looking for a specific information and not a document. In the literature, great efforts which have been made to build a reliable QA system are reported for most languages. However, attempts to investigate the implementation of a QA system oriented to the Arabic language are scarce. In this paper, we present an early version of a QA system oriented to the Arabic language and which we hope to release soon. We give a generic description of the different parts of the system and also draw results of the already implemented parts.**

I. Introduction

Nowadays, the Web has become the main source of information as nearly all kind of data (digital libraries, newspapers collections, etc.) stored in electronic format. The data available is likely to satisfy most requests, nevertheless without the appropriate search facilities, the great amount of retrieved information is practically useless. In fact, the mechanisms developed up to now in *Information Retrieval (IR)* - those used for instance, by *search engines* such as *Google*[1], *Yahoo*[2] or *MSN*[3] - allow a user only to retrieve the relevant documents which (partially) match a given query [1]. It is the user's task to look for the information within the relevant documents themselves once they are retrieved.

In recent years, the combination of the web growth and the explosive demand for better information access has motivated the interest in *QA* systems [2]. The goal of a QA system is to provide inexperienced users with a flexible access to information allowing them for writing a query in natural language and obtaining not the documents which contain the answer, but the concise answer itself. For instance, fgiven the user's question:

Who is the President of the Syrian Arab Republic?

we do not want a search engine to retrieve relevant *snippets* (such as the one below) containing a link to a document to extract the information from; we would prefer instead a QA system which could instantly return the exact answer: *Bashar Al-Assad*.

[1] www.**google**.com/
[2] www.**yahoo**.com/
[3] www.**msn**.com/

Syria - Wikipedia, the free encyclopedia
Al- umhūriyyah al-Arabiyyah as-Sūriyyah **Syrian Arab Republic ...** In May 2004, the **President** claimed that **Syria** had not met these conditions and implemented **...**
en.wikipedia.org/wiki/Syria - 174k - Cached - Similar pages

In comparison with other languages, not many are the NLP tools and resources in general (corpora, gazetteers, etc) which are available for Arabic. This is specially true for the QA task. In the international Text Retrieval Conference (TREC[4]) and Cross-Language Evaluation Forum (CLEF[5]) competitions, there is not a QA task at the moment which includes the use of the Arabic language and so far, only two Arabic Cross-Language IR (CLIR) tasks were organised in the TREC 2001 and 2002 competitions (but none proposed a QA task in Arabic). Therefore, QA systems are often developed for English as the target language and because English is the language of the majority of the documents on the Web. In [3] preliminary Cross-Language Question Answering (CLQA) experiments were carried out in order to allow for querying the system in the Arabic language and translating each question into English. The main aim was to investigate the effect of using a translator (from Arabic into English) in a QA system.

Two *Arabic* QA systems - AQAS [4] and QARAB [5] – have been developed so far. Unfortunately, the AQAS system is knowledge-based and, therefore, extracts answers only from structured data and not from raw text (non structured text written in natural language); moreover, no results are presented in the paper. The authors of the QARAB system instead report a *precision* (ratio of number of correctly answered questions / number of answered questions) and a *recall* (number of correctly answered questions / number of questions) of 97.3%. The evaluation was done directly by the four native Arabic speakers (university graduates) who presented the 113 questions to the system and judged *themselves* the correctness of the answers. These (possibly biased) results are not reliable because such accuracy was not not achieved in any other language in the QA state-of-the-art. For this reason, we believe that first of all a test-bed of questions in Arabic should be provided in order to allow a comparison between different QA systems (see sub-section III.2.).

The rest of this paper is structured as follows: Section two describes the generic architecture of our Arabic QA system. The third Section is devoted to the description of the components which we have already implemented: the Named Entity Recognition system used in our QA system and the Passage Retrieval module. In Section four we describe the components not implemented yet. Finally, we draw some conclusions and discuss the further work to be done to improve the elements who were implemented as well as those to be implemented to complete our QA system.

II. The generic architecture of our future Arabic QA system

The generic architecture of a QA system is illustrated in Figure 1. It is composed of three main elements/parts :

[4] http://**trec**.nist.gov/
[5] http://www.**clef**-campaign.org/

(i) the *Question Analysis* module which processes the question in order to obtain some useful information about the type of answer we are looking for and extracts the question key words and named entities;

(ii) the *Passage Retrieval* (PR) module which has to process the documents in order to retrieve those passages with the highest probability of containing the answer;

(iii) and finally, the *Answer Extraction* (AE) module which extracts the answer from the retrieved passages taking into account the constraints of the Question Analysis module.

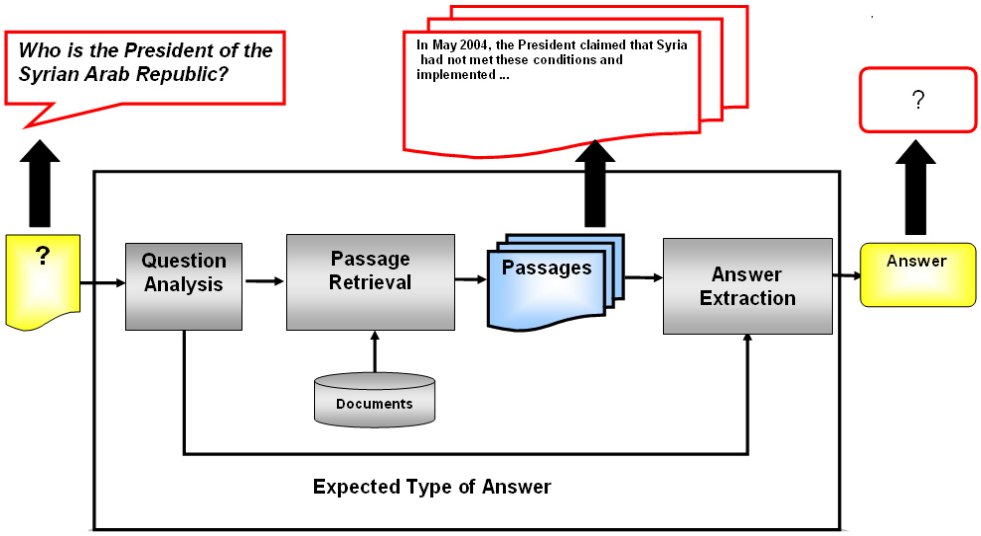

Fig. 1. The generic architecture of a Question Answering system.

It is also very important to point out that a Named Entities Recognition (NER) is required as a tool for almost all the QA system components. Those NER systems allow to extract proper nouns as well as temporal and numeric expressions from raw text. In our case we used a NER system implemented by ourselves [6] especially for the Arabic open-domain texts. We have considered four proper names categories:

(i) Organization: named corporate, governmental, or other organizational entity;
(ii) Location: name of politically or geographically defined location;
(iii) Person: named person or family;
(iv) Miscellaneous: other proper names.

For more details about our NER system see Section three for more details about our NER system.

III. Implemented components

III.1.ANERsys: An Arabic NER system

In order to determine the named entities in the question and in the relevant passages returned by the PR module , there is a need for a reliable NER system especially designed for the Arabic language. Nowadays, there are no Arabic NER

systems available for free use. To tackle this problem we had to implement our own NER system. The Arabic language has the peculiarity of not having capital letters, this makes the NER task even more challenging but not impossible!
The choice of the approach to use in order to implement the NER system was based on the study of the existing language-independent NER systems. The best participations in the shared task (which concerned language-independent NER systems) of the CONLL 2002[6] adopted the Maximum Entropy (ME) model [7][8][9][10][11]. Moreover, in [11] a comparison was made between the Hidden Markov Models and the ME based NER systems where the ME one showed better results. For our Arabic NER system we chose the ME approach because it proved to be efficient for the Arabic language too (all the details about the implementation are given in [6]).

In order to train, evaluate and boost the system we had also to develop a corpus (*ANERcorp:* 150,000 tagged tokens) as well as few gazetteers (*ANERgazet*) which are now freely available[7] to the research community interested in working with texts in Arabic. Table 1 shows the final results obtained with this system.

	Precision	Recall	F-measure
Location	82.17%	78.42%	80.25%
Misc	61.54%	32.65%	42.67%
Organisation	45.16%	31.04%	36.79%
Person	54.21%	41.01%	46.69%
Overall	**63.21%**	**49.04%**	**55.23%**

Table 1. Results of our Arabic NER system

III.2.Arabic-JIRS: A PR system adapted to the Arabic language

The PR module is no doubt the core part of a QA system because it is evident that if swe do not succeed in retrieving the relevant passages containing the answer, it will be impossible to extract the answer in the following module. In our implementation we adapted to Arabic the language-independent JIRS passage retrieval system[8] which was developed in our Natural Language Engineering Lab. research team. JIRS is a QA-oriented system which relies on an n-gram model in order to index the documents. For the retrieval of the relevant passages, it uses a Distance Density model which gives more weight to the passages where the question structures appear nearer to each other, [12]. Figure 2 illustrates the architecture of JIRS.

In order to omit affixes and, therefore, to overcome the data sparseness problem of Arabic texts, a light stemmer[9] was used. To avoid a manual estimation for the JIRS performance on Arabic text, a test-bed[10] for Arabic QA (with 11,000 documents of the Arabic *Wikipedia*[11], 200 questions and a set of correct answers) was recently built

[6] http://www.cnts.ua.ac.be/conll2002/ner/
[7] http://www.dsic.upv.es/~ybenajiba
[8] Available at: http://jirs.dsic.upv.es
[9] Available at: http://www.glue.umd.edu/kareem/research/
[10] Available at: http://www.dsic.upv.es/~ybenajiba
[11] http://ar.wikipedia.org

following the guideline of CLEF. The preliminary results, [8], seem to be very promising and a *coverage* measure of passages containing the right answer of 69% was obtained. More details about the adaptation of the JIRS to the Arabic language can be found in [13].

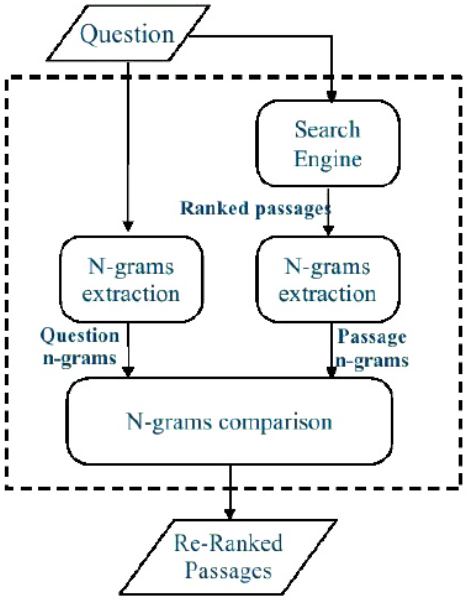

Fig. 2. The JIRS generic Architecture

IV. Components to be implemented

In this Section we will describe the remaining components that we propose to develop further.

IV.1.Question Analysis module:

This module is required for the:

1- Extraction of the question keywords: this operation consists of eliminating the stopwords (words which are useless for the question analysis task) appearing in the question;

2- Extraction of the question named entities: for this operation we will use the NER system we have built (see Section III.1.)

3- Classification of question: finally, we will have to give the question type which is very important for the answer extraction. In order to constantly remain within the CLEF guidelines we have chosen the following question types:

(i) Name: Acronym, Person, Title, Location (Country, City, Geographical);
(ii) Definition;
(iii) Date: Day, Month, Weekday;
(iv) Quantity: Money, Dimension, Age.

IV.2. Answer Extraction module

The AE module is the component that selects the candidate answers from the relevant passage returned by the PR module taking in consideration the question type, i.e. the expected type of answer. We are planning to build an AE module composed of two main parts:

(i) AE for factoid questions, which is only invoqued in case the answer to return is a proper name, a quantity or a date;
(ii) AE for other questions: it will be especially concerning more complicated questions and will use semantic information to extract the candidate answers.

The first part of the AE module will employ the machine learning technique described in [14] and will perform in three main steps:
(i) Tag the named entities of the relevant passage;
(ii) Pre-selection of the candidate answers based only on the type of the expected answer;
(iii) Selection of the final list of candidate answer based on a pattern filtering (Figure 3 gives an illustrating example).

With respect to the second part, we have not figured out how it could exactly perform because we have first to implement some of the required semantic tools (Thesaurus, Corpora, etc.)

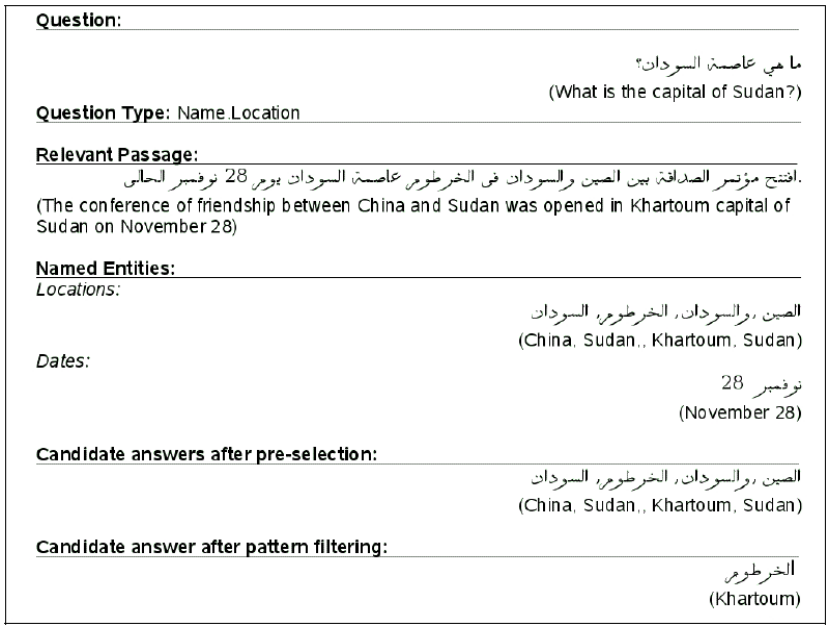

Fig. 3. Illustrating example of the future AE module's steps

V. Conclusion and Future Work

In this paper we described the generic architecture of the Arabic QA system that we are implementing. We also showed the already implemented NER system, which is required for most of the system components. A ME approach was employed obtaining

good results which we plan to improve in the future to get a higher accuracy. The PR module, which is the core component of the system, was implemented by adapting an existing language-independant PR system which enabled a coverage up to 69%. Our next objective is to test our PR module on a bigger corpus as we believe that we can obtain higher coverage with a corpus where the information is more redundant.

At the moment of writing this paper, some efforts have been making to build the Question Analysis and the Answer Extraction modules in order to complete the implementation of the Arabic QA system,

Aknowledgements

The research work of the first of the authors mentioned above was partially supported by MAEC - AECI. We would like to thank the AECI-PCI Spain-Morocco A/7067/06 and MCyT TIN2006-15265-C06-04 research project for partially funding this work.

References

[1] Baeza, R., and Ribeiro, B., *Modern Information Retrieval.* ACM Press, New York, Addison-Wesley, 1999.

[2] Burger, J. et alii. *Issues, tasks, and program structures to roadmap research in question & answering (q&a)*, in: NIST, 2002.

[3] Rosso, P., Lyhyaoui, A., Peñarrubia, J., Montes y Gómez, M., Benajiba, Y., Raissouni, N., *Arabic-English Question Answering.* In: Proc. Symposium on Information Communication Technologies Int., Tetuan, Morocco, 2005.

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

[5] Hammou, B., Abu-salem, H., Lytinen, S., Evens., M., QARAB: *A Question answering system to support the ARABic language.* In: Proc. of the workshop on Computational approaches to Semitic languages, ACL, pages 55-65, Philadelphia, 2002.

[6] Benajiba, Y., Rosso, P., Benedí, J. M., ANERsys: *An Arabic Named Entity Recognition System based on Maximum Entropy.* 8th Int. Conf. on Intelligent Text Processing and Computational Linguistics (CICLing 2007). LNCS, Vol. 4394, Springer-Verlag, pp.143-153 (to be published).

[7] O. Bender, F. J. Och, H. Ney, *Maximum Entropy Models For Named Entity Recognition.* In Proceedings of CoNLL-2003. Edmonton, Canada, 2003.

[8] L. Chieu Hai, T. Ng Hwee, *Named Entity Recognition with a Maximum Entropy Approach*. In Proceedings of CoNLL-2003. Edmonton, Canada, 2003.

[9] JR. Curran, and S. Clark, *Language Independent NER using a Maximum Entropy Tagger*. In Proceedings of CoNLL-2003. Edmonton, Canada, 2003.

[10] S. Cucerzan and D. Yarowsky, *Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence*. In Proceedings, 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora, pp.90–99.

[11] R. Malouf, *Markov Models for Language-Independent Named Entity Recognition*. In Proceedings of CoNLL-2003. Edmonton, Canada, 2003.

[12] Gómez, J., Montes, M., Sanchís, E., Villaseñor, L., Rosso, P. Language *Independent Passage Retrieval for Question Answering*. LNAI-Springer Verlag, vol. 3789, 2005.

[13 Benajiba, Y., Gómez, J. M., Rosso, P., *Adapting JIRS Passage Retrieval System to the Arabic Language.* 8th Int. Conf. on Intelligent Text Processing and Computational Linguistics (CICLing 2007). LNCS, Vol. 4394, Springer-Verlag, (to be published).

[14] Téllez, A., Montes, M., Villaseñor, L., *A Machine Learning Approach for Information Extraction.* In: Proc. 6th Int. Conf. on Intelligent Text Processing and Computational Linguistics (CICLing), Springer-Verlag, vol. 3406.