Arabic QA4MRE at CLEF 2012: Arabic Question Answering for Machine Reading Evaluation

Omar Trigui[1], Lamia Hadrich Belguith[1], Paolo Rosso[2], Hichem Ben Amor[1], Bilel Gafsaoui[1]
[1]University of Sfax, ANLP Research Group- MIRACL Laboratory, Tunisia
[2] Natural Language Engineering Lab - ELiRF,Universitat Politècnica de València, Spain
omar.trigui@fsegs.rnu.tn, l.belguith@fsegs.rnu.tn, prosso@dsic.upv.es, hichem.ben.amor.fsegs@gmail.com, gafsaouibilel@gmail.com

Abstract. This paper presents the work carried out at ANLP Research Group for the CLEF-QA4MRE 2012 competition. This year, the Arabic language was introduced for the first time on QA4MRE lab at CLEF whose intention was to ask questions which require a deep knowledge of individual short texts and in which systems were required to choose one answer from multiple answer choices, by analyzing the corresponding test document in conjunction with background collections. In our participation, we have proposed an approach which can answer questions with multiple answer choices from short Arabic texts. This approach is constituted essentially of shallow information retrieval methods. The evaluation results of the running submitted has given the following scores: accuracy calculated overall all questions is 0.19 (i.e., 31 correct questions answered correctly among 160), while overall c@1 measure is also 0.19. The overall results obtained are not enough satisfactory comparing to the top works realized last year in QA4MRE lab. But as a first step at the roadmap of the evolution of the QA to Machine Reading (MR) systems in Arabic language and with the lack of researches investigated in the MR and deep knowledge reasoning in Arabic language, it is an encouraging step. Our proposed approach with its shallow criterion has succeeded to obtain the goal fixed at the beginning which is: select answers to questions from short texts without required enough external knowledge and complex inference.

Keywords: Machine Reading, Reading Comprehension, Knowledge Reasoning, Arabic Language.

1. Introduction

The QA4MRE lab this year was aimed to evaluate machine reading systems which require a deeper level of text understanding to answer questions with multiple answer choices in a set of seven languages[1]. The way proposed to assess the understanding of a text is the ability to answer a set of questions about it. This evaluation manner is similar to reading comprehension tests designed to evaluate how well a human reader has understood a text.

This paper is structured into 6 sections. Section 2 presents a short overall of the state of the arts of Machine Reading Evaluation. Section 3 details our proposal approach to deal with QA4MRE lab in Arabic language. Section 4 presents the experiment carried out. Section 5 discusses the obtained results, and finally, Section 6 presents a general conclusion and perspective.

2. Related works

The important beginning of the researches on Machine Reading (MR) were on 'Reading comprehension tests as evaluation for computer-based language understanding systems' workshop organized in 2000[2]. Then, separate researches were done on MR such as [1] until last year where in the framework of CLEF the first version of the Question Answering for Machine Reading Evaluation lab (QA4MRE lab) in five European language versions has been organized. Mainly, the approaches proposed by different researchers show how to adapt QA systems to MR systems. This year CLEF organizes QA4MRE lab with two new languages: Arabic and Bulgarian with more difficult task than in 2011.

[1] http://celct.fbk.eu/QA4MRE/index.php?page=Pages/mainTask.html
[2] http://portalparts.acm.org/1120000/1117595/fm/frontmatter.pdf?ip=41.228.226.199&CFID=142563908&CFTOKEN=51195488

3. Proposed approach

Our proposed approach is constituted of a shallow process to understand texts based on Information Retrieval (IR) and it requires inferring from texts. Four steps are involved in the shallow method after the preprocess of the corpus based on the anaphoric resolution. The first step is about the question analysis where the stop words in the question were removed and the rest of words are saved as the question focus. The second step is the research of passages containing the focus question. In the third step, the passages collected are aligned with the multiple answer choices for the respective question. The answer which is included at these passages is selected as the correct answer from the multiple answer choices. In case, where there is no answer included in passages. We introduce a list of inference constituted of pair of words generated from the background collection according to a list of inference rules. In the alignment process, any word from the answer option that does not exist in the passage is replaced by its respective inference word. If after using the inference there is not an answer included in the collected passages, the question is considered as unanswered.

This approach is proposed for dealing exactly with non complex question types such as the factoid question type, and questions whose answers are selected from their respective single short texts. Table1 illustrates an example of these questions introduced in the data set of QA4MRE lab 2012. This part of questions is similar to the part of 76 questions (63%) of the 120 questions used in QA4MRE lab in CLEF 2011 and that do not require extra information from the background collection in order to be answered. They require just information presented in a paragraph or a sentence [2].

Table 1. An example of a factoid question related to Aids topic of QA4MRE lab CLEF of 2012. The correct answer is in bold.

A question in Arabic language	متى بدأت الهند تعترف ببراءات اختراع المستحضرات الصيدلانية؟
The translation in English language	When India began to recognize the patented pharmaceutical products?
The given multiple answer choice	1991
	2007
	1982
	2005
	1997

4. Experimentation

The proposed approach was implemented in a system following the architecture shown in Figure 1. The module ‘preprocess of the corpus’ where the anaphoric resolution and the construction of inference list by topic according to inference rules were not integrated. The evaluation of our system was carried out using the data given by the CLEF organization. The test set is composed of 16 test documents, 160 questions with a set of five answer choices per question. The test documents are related to 4 topics (i.e. “Aids”, “Climate change”, “Music and Society”, and “Alzheimer”). Each topic includes 4 reading tests and each reading test is composed of one single document with 10 questions related to it. A background collection consisted of un-annotated documents related to the 4 topics in 7 languages are given to the participants to acquire the reading capabilities and the knowledge needed to fill in the gaps required to answer a test on the topic. Our system was required to answer these 160 questions by selecting one answer from the five alternative answers. There is always one and only one correct answer for each question. We have the option to leave a question unanswered if there is no confidence about the correctness of its response.

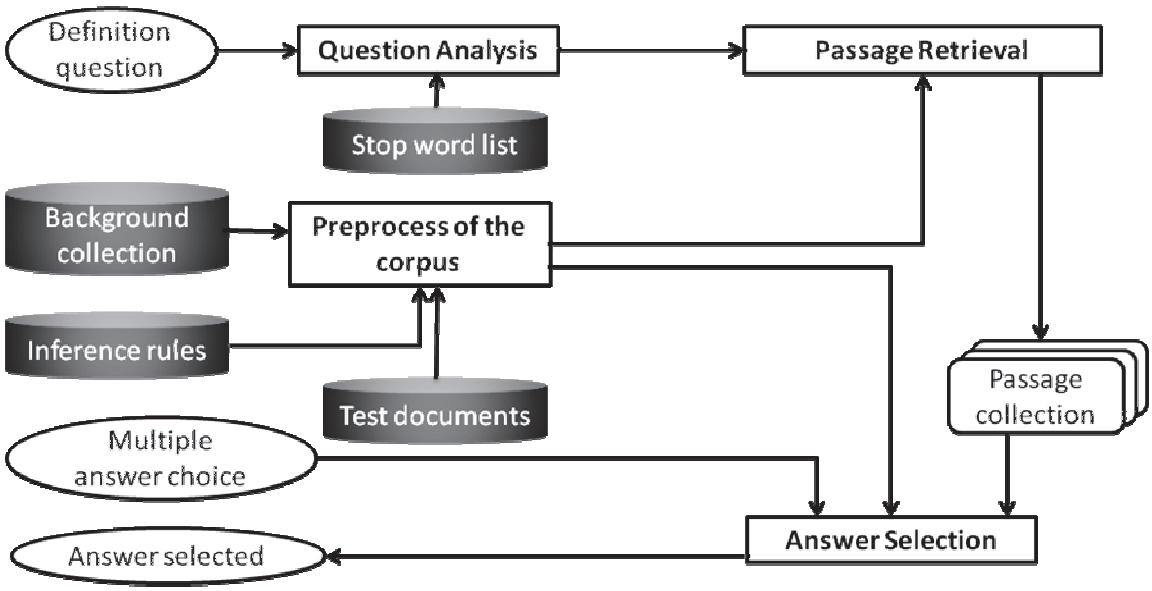

Figure 1. The architecture of the system used at QA4MRE lab 2012

The measure c@1 is applied to encourage systems to leave some questions unanswered in order to reduce the amount of incorrect answers [3].

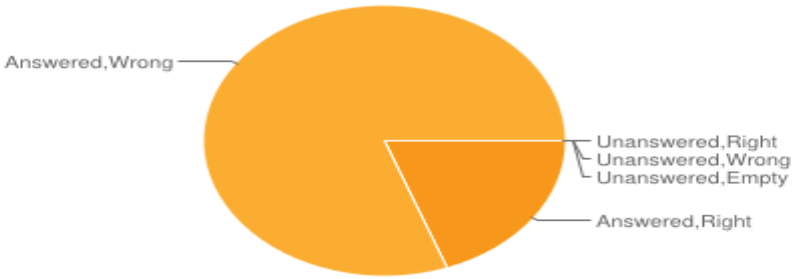

Figure 2. The questions distribution according to their answers

The results of the evaluation measures are as follows: the overall accuracy is equal to 0.19 (i.e. we have succeeded to answer correctly to 31 questions from 160 questions), and C@1 is equal also 0.19

Overall accuracy = (nr / n)
C@1 = (nr + nu * (nr/n)) / n

where:
nr: is the number of correctly answered questions
nu: is the number of unanswered questions
n: is the total number of questions

Table 2. The number of the answered and unanswered questions by our system

Number of questions answered	**160**
ANSWERED with RIGHT candidate answer	31
ANSWERED with WRONG candidate answer	129
Number of questions UNANSWERED	**0**

Table 3. Average and overall c@1 measures per topic of our system

Topics	**Overall c@1 per topic**	**Average per topic**
Aids	0.23	0.23
Climate change	0.25	0.25
Music and Society	0.15	0.15
Alzheimer	0.15	0.15

Table 2 and Figure 2 illustrate the part of the questions answered correctly and those answered wrongly. While Table 3 shows the overall c@1 and average per topic. The best measures are realized by "climate change" topic because it contains a big part of its questions in simple type (i.e. factoid).

5. Discussion

The overall results obtained are not comparable to the top performance obtained last year for the English language. Nevertheless as a first step at the roadmap of the evolution of the QA to MR evaluation systems in Arabic language and with the lack of researches investigated in deep knowledge reasoning in Arabic language, it could be considered as an encouraging step. Our proposed approach with its shallow criterion has succeeded to obtain the goal fixed at the beginning which is: selecting answers to questions from short texts without required external knowledge and complex inference.

6. Conclusion

The QA4MRE lab was focused this year on the evaluation of Machine Reading systems. The goal behind it was to push researches towards a deeper understanding of a single text using inference deduced from document collection. We have participated at this QA4MRE lab which has included Arabic language for the first time. In our work, we have proposed an approach which did not require a deep reasoning and inference. We have succeeded to a certain degree to obtain an overall accuracy of 0.19. We plan in the future to improve this result by investigating further in MR research deep knowledge reasoning in Arabic language.

Acknowledgments. The European Commission as part of the WIQ-EI IRSES-Project (grant no. 269180) within the FP 7 Marie Curie People Framework has partially funded the work of the third author. His work was carried out also in the framework of the MICINN Text-Enterprise (TIN2009-13391-C04-03) research project and the Microcluster VLC/Campus (International Campus of Excellence) on Multimodal Intelligent Systems.

References:

1. Ben Wellner , Lisa Ferro , Warren Greiff , Lynette Hirschman, Reading comprehension tests for computer-based understanding evaluation, Natural Language Engineering, v.12 n.4, p.305-334, December 2006.

2. Anselmo Peñas, Eduard H. Hovy, Pamela Forner, Álvaro Rodrigo, Richard F. E. Sutcliffe, Corina Forascu, Caroline Sporleder: Overview of QA4MRE at CLEF 2011: Question Answering for Machine Reading Evaluation. CLEF (Notebook Papers/Labs/Workshop) 2011.

3. Anselmo Peñas and Alvaro Rodrigo. A Simple Measure to Assess Non-response. In Proceedings of 49th Annual Meeting of the Association for Computational Linguistics - Human Language Technologies (ACL-HLT 2011). Portland. Oregon. USA. June 19-24. 2011.