Question-Answering and E-Health

An advanced perspective in Specialized Information Retrieval

María-Dolores Olvera-Lobo
Dept. of Information and Communication, University of Granada, UGR, Granada, Spain & Unidad Asociada Grupo SCIMAGO, CSIC, Madrid, Spain
molvera@ugr.es

Juncal Gutiérrez-Artacho
Department of Translation and Interpreting
University of Granada, UGR
Granada, Spain
juncalgutierrez@ugr.es

***Abstract*— In the field of Information Retrieval (IR), new approaches have emerged in order to offer more efficient data/information. This is the case with Question Answering (QA) systems that try to offer precise and understandable answers to factual questions. We have analyzed the state of Biomedical QA systems, we have evaluated the performance of those systems on the Web, and we present new trends in this emerging area.**

***Keywords*— Question Answering Systems, Information Retrieval, Restricted-domain QA Systems, Biomedical QA System, e-Health**

I. Introduction

The Web and its subsequent expansion have provided the general public with access to enormous volumes of information, offering unquestionable benefits. QA systems constitute an alternative to IR systems, as they aim at automatically finding concise and understandable answers to factual questions, rather than just offering a list of documents that are related to the search.

Within this framework, in previous studies [1-5] we have evaluated the quality and effectiveness of tools for the open and restricted-QA systems, by analyzing and comparing the results with the aim of contributing to the development of this emerging subarea of the IR and providing a new source to retrieve medical information for physicians and users. In this paper, we offer an overview of the state of QA systems in the field of health, with special emphasis on the Web context. In addition, we summarize our recent findings and we present new trends and applications in biomedical QA systems.

II. State of the Art

The exponential growth in the volume of publications in the biomedical domain has made it impossible for an individual to keep pace with advances. Even though evidence-based medicine has gained wide acceptance, physicians are unable to access the relevant information in the required time, leaving most of the questions unanswered. This accentuates the need for fast and accurate biomedical QA systems [6]. These QA systems can exploit deeper text analysis/ processing, by taking advantage of domain-specific formatting and style conventions as well as domain-dependent terminology. Athenikos and Hans [7] summarize the main characteristic features of QA in the biomedical domain as large-sized corpora; highly complex domain-specific terminology, domain-specific lexical, terminological, and ontological resources, tools and methods for exploiting the semantic information embedded in the above resources and domain-specific format and typology of questions.

III. Evaluations in Biomedical QA Systems

Previous studies to validate the usefulness of biomedical QA systems have revealed serious problems in the biomedical QA process [7-8], since many clinical questions have gone unanswered. For this reason, we carried out several studies aimed at designing and developing a methodological approach for the evaluation of QA systems in the context of the Web. The question collection was composed of a sample of medical questions that made it possible to evaluate the performance and the sources of information used by the analyzed QA systems. After asking questions in the different QA systems, a group of medical professionals and students of Medical Studies evaluated the answers as incorrect, inexact or correct. The evaluation measures applied to assess the answers retrieved were: *Mean Reciprocal Rank* (MRR), *Total Reciprocal Rank* (TRR), *First Hit Success* (FHS), *the average precision* (AveP), *precision* and *recall.*

A. Results: Open- vs. Restricted- Domain QA Systems

We have compared open- vs. restricted-domain QA systems. After asking the 150 questions in the four QA systems (START, QuALiM, MedQA and HONqa), the first five answers from each system were analyzed. The average of answers retrieved by each QA systems was very significant.

	Total answer	*Average answers*	*% correct answers*	*% incorrect answers*	*% inexact answers*
HONqa	6635	44.23	47.24	44.76	7.99
QuALiM	441	3	40.88	44.33	14.78
MedQA	802	5.34	46.66	34.75	18.58
START	236	1.6	70.08	11.54	18.38

a. Answers in the evaluated QA systems

The correct answers are present to the greatest degree in START (70.08%). In the two restricted-domain QA systems, this average decreases – MedQA (46.66%) and HONqa (47.24%) – and QuALim is the most deficient, with 40.88% of answers correct. The value obtained by applying the MRR and FHS evaluation measures indicates that MedQA best ranks answers, as the first correct answer appears at the top of the list of results. As can be observed, none of the applied metrics present very high values, a situation which has clearly been influenced by the high standards set for an answer to be evaluated as correct.

	MRR	*TRR*	*FHS*	*Precision*	*AveP*	*Recall*
HONqa	0.75	1.15	0.55	0.47	0.53	0.46
QuALiM	0.65	0.77	0.59	0.40	0.43	0.22

MedQA	0.87	1.29	0.76	0.46	0.65	0.44
START	0.67	0.81	0.64	0.70	0.88	0.22

b. Evaluation measures for QA systems

B. Results: Multilingual Biomedical QA Systems

We asked the questions in the multilingual QA system, HONqa, and the first five answers in each of these systems were then analyzed. Although the mean of the answers retrieved by the system in the three languages approached and in some cases exceeded 30, only the first answers offered were considered. The average of the total answers retrieved by the system was 47.46 in the case of English, 27.36 for French, and 25.03 for Italian.

	Total answers	*Average answers*	*% Correct answers*	*% Inexact answers*	*% Incorrect answers*
English	5695	589	48.73	11.37	39.9
French	3283	573	9.07	21.64	69.28
Italian	3123	585	5.47	11.63	82.9

c. Answers retrieved by HONqa in the three languages

The correct answers were present in greater measure in the English version of the system, which properly responded to more than 48% of the cases, whereas French offered a low rate of 9.07% and Italian provided only 5.47%. The MRR value for the responses offered in the three languages reflect the above comments. In relation to the TRR measure, it was found that, except for English the results did not substantially improve. FHS is an important measure, as the users often tend to focus on the first response retrieved, skipping the rest. It was found that more than 50% of the answers offered in English (0.575) provided an initial correct answer while the other cases were not encouraging (0.12 in French and 0.06 in Italian).

	MRR	*TRR*	*FHS*	*Precision*	*AveP*	*Recall*
English	0.76	1.55	0.575	0.55	0.65	0.59
French	0.19	0.27	0.12	0.10	0.31	0.17
Italian	0.13	0.15	0.06	0.05	0.16	0.13

e. Evaluation measures

The precision value is closely related to the rest of the measures discussed above. The small number of correct answers in some cases made the recall values of the QA system very low, except in the case of English.

IV. New Trends in QA Systems: ibm Watson and Wolfram alpha

Although QA systems have yet to receive the recognition they deserve and their availability is still very low for final users, currently there are new approaches that show the way for new QA systems. This is the case with IBM Watson and with Wolfram Alpha.

IBM Watson is a QA system built to compete at the human champion level in real time on the American TV quiz show, Jeopardy. The system is workload optimized, integrating massively parallel POWER processors and being built based on IBM's DeepQA technology [9], which it uses to generate hypotheses, gather massive evidence, and analyze data.

Wolfram|Alpha introduces a fundamentally new way to get knowledge and answers—not by searching the Web, but by doing dynamic computations based on a vast collection of built-in data, algorithms, and methods. This QA system collects and curates all objective data, implements every known model, method, and algorithm, and makes it possible to compute whatever can be computed about anything.

V. Conclusions

Users are more demanding of retrieved information, both regarding quality and quantity, and regarding response times. Because of this, QA systems could be one of the future information retrieval systems on the Web, since they attempt to meet the needs and demands of current users. The analysis of the results from asking questions in biomedical QA systems has enabled the evaluation of its performance in the retrieval of multilingual information by applying specific measures and analyzing the information sources used for each language. Despite the restrictions that these systems show, the study indicates that this QA system is valid and useful for the retrieval of medical definition information, mainly in English, although it is not yet the most recommended resource to gather multi-lingual information in a quick and precise way.

As we can see with the case of Watson and Wolfram|Alpha, QA systems have been extended in recent years to explore critical new scientific and practical dimensions. Future research may explore what kinds of questions can be asked and answered about social media, including sentiment analysis. It remains necessary to deepen the interactive design of these systems and enable true feedback between questions and answers, so that users communicate with the system in a conversational manner.

References

[1] M.D. Olvera-Lobo, and J. Gutiérrez-Artacho. “Evaluation of Open- vs. Restricted- Domain Question Answering Systems in the Biomedical Field”. Journal of Information Science, 37, vol. 2, 2011, pp. 152-162

[2] M.D. Olvera-Lobo, and J. Gutiérrez-Artacho. “Question-Answering Systems as Efficient Sources of Terminological Information: Evaluation”. Health Information and Library Journal, 27, vol. 4, 2010, pp. 268 – 274.

[3] M.D. Olvera-Lobo, and J. Gutiérrez-Artacho. “Multilingual Question-Answering System in biomedical domain on the Web: an evaluation”. LNCS, vol. 6941, 2011

[4] M.D. Olvera-Lobo, and J. Gutiérrez-Artacho. “Language resources used in multi-lingual Question Answering Systems”. Online Information Preview. 35, vo. 4, 2011, pp. 543 – 557

[5] M.D. Olvera-Lobo, and J. Gutiérrez-Artacho. “Evaluación del rendimiento de los sistemas de búsqueda de respuestas de dominio general”. Revista Española de Documentación Científica, 36, vol. 2, e009

[6] G.O. Sing, C. Ardil, W. Wong, and S. Sahib. “Response Quality Evaluation in Heterogeneous Question Answering System: A Black-box Approach”. Proceedings of World Academy of Science, Engineering and Technology, 9, Lisbon, 2005.

[7] S.J. Athenikos, and H. Han. “Biomedical question answering: A survey”. Computer Methods and Programs in Biomedicine 99, vol. 1, 2010, pp. 1-24

[8] J.W. Ely, J. Osheroff, P.N. Gorman, M.H. Ebell, M.L. Chambliss, E.A. Pifer, and P.Z. Stavri. “A taxonomy of generic clinical questions: classification study”, BMJ, vol. 321, 2000, pp. 429–432

[9] Is Watson the smartest machine on earth?, Computer Science and Electrical Engineering Department, UMBC, 2011