Question Answering Track Evaluation in TREC, CLEF and NTCIR

María-Dolores Olvera-Lobo[1,2] and Juncal Gutiérrez-Artacho[3]

[1] University of Granada, Department of Information and Communication,
Colegio Máximo de Cartuja, Campus Cartuja s/n, 18071, Granada, Spain
[2] CSIC, Unidad Asociada Grupo SCImago, Madrid, Spain
[3] University of Granada, Department of Translation and Interpreting,
Faculty of Translation and Interpreting, C/ Buensuceso, 11, 18003, Granada, Spain
{molvera,juncalgutierrez}@ugr.es

Abstract. Question Answering (QA) Systems are put forward as a real alternative to Information Retrieval systems as they provide the user with a fast and comprehensible answer to his or her information need. It has been 15 years since TREC introduced the first QA track. The principal campaigns in the evaluation of Information Retrieval have been specific tracks focusing on the development and evaluation of this type of system. This study is a brief review of the TREC, CLEF and NTCIR Conferences from the QA perspective. We present a historical overview of 15 years of QA evaluation tracks using the method of systematic review. We have examined identified the different tasks or specific labs created in each QA track, the types of evaluation question used, as well as the evaluation measures used in the different competitions analyzed. Of the conferences, it is CLEF that has applied the greater variety of types of test question (factoid, definition, list, causal, yes/no, amongst others). NTCIR, held on 13 occasions, is the conference which has made use of a greater number of different evaluation measures. Accuracy, precision and recall have been the three most used evaluation measures in the three campaigns.

Keywords: Question Answering Systems, Evaluation metrics, QA Tracks, TREC, CLEF, NTCIR.

1 Introduction

Frequently, a keyword query entered into a web search tool (search engine or meta-search engine) to satisfy a user's information need, provides too many result pages – many of which are useless or irrelevant to the user. In effect, modern Information Retrieval (IR) systems allow us to locate documents that might have the associated information, but the majority of them leave it to the user to extract the useful information from an ordered list [1]. In contrast to the IR scenario, a Question Answering (QA) system processes questions formulated into Natural Language (NL) instead of keyword based queries, and retrieves answers instead of documents [2]. Therefore, the usefulness of these types of systems for quickly and effectively finding specialized information has been widely recognized [3,4].

© Springer International Publishing Switzerland 2015

Á. Rocha et al. (eds.), *New Contributions in Information Systems and Technologies*,
Advances in Intelligent Systems and Computing 353, DOI: 10.1007/978-3-319-16486-1_2

The aim of QA systems is to find precise and correct answers for users' questions. QA systems perform three basic actions: question analysis, passage retrieval, and answer extraction. In the question analysis process, the system locates words that offer some clues about the type of answer we are looking for question words, nouns, adjectives or verbs. The system also seeks text elements and classifies them into categories such as the names of people, organizations, locations, expressions of time, quantities, monetary values, etc. Later, it processes the documents in order to retrieve those parts of text with the highest probability of containing the answer. Finally, it extracts the definitive answer.

The evaluation of QA systems is a major research area that needs attention, especially with the emergence of domain-oriented question answering systems based on natural language understanding and reasoning [5]. Although we find different analyses referring to the evaluation of QA systems, such as those focusing on the evaluation of QA systems on the Web [6-11], or in some of the international evaluation forums [2, 12], up to now there has been no broad analysis of the use of evaluation measures of these systems in the main international forums.

Every year the organizers of the major conferences on this issue propose a particular objective for each specific QA track. Then, they choose an appropriate evaluation method which involves determining specific features of the collections, and selecting the measures for assessing the performance of participating systems [2]. This study is primarily intended as a general-purpose review, and aims to identify and analyze briefly several aspects of the different QA tracks in TREC, CLEF and NTCIR. We have examined the reports and proceedings of those conferences from when they began up to but not including 2015.

2 Question Answering in Information Retrieval Evaluation Campaigns

2.1 Text REtrieval Conference (TREC)

TREC introduced the first QA track in TREC-8 (1999). During the following 8 conferences -the QA track ran out of in 2007-, the aim of the TREC QA campaigns was to assess the capability of systems to return exact answers to open-domain English questions. The QA track at TREC was the first attempt in IR to stress the importance on these systems [2]. TREC workshops consisted of a set of areas of focus in which particular retrieval tasks are defined. The tasks in the track have evolved over the years to focus research on particular aspects of the problem deemed important to improving the state-of-the-art [13]. Table 1 shows the different tasks, which respond to research concerns and developments in this field, performed between 1999 and 2007 focusing on QA systems.

Table 1. Tasks in QA Track in TREC

Task	Characteristics	No. Times performed
Main task	Give the most up-to-date answer supported by the document collection	9
List task	Require systems to assemble an answer from information located in multiple documents	2
Context task	Test systems' ability to track discourse objects (context) through a short series of questions	1
Document ranking	Build infrastructure that would allow a closer examination of the role document retrieval techniques play in supporting QA technology	1
Relationship task	Give to systems topic statements that ended with a question asking for evidence for a particular relationship	1
CiQA (complex task)	Promote the development of interactive systems capable of addressing complex information needs	2
Passages	Find a small chunk of text that contains the exact-phrase answer of a given question from a large document collection	1

2.2 Conference and Labs Evaluation Forum (CLEF)

The first QA Track at CLEF took place in 2003. The systems were fed with a set of questions and were asked to return one or more exact answers per question –where *exact* means that neither more nor less than the information required is returned. The answer needed to be supported by the identification of the document in which the exact answer was found, and depending on the year, also by portion(s) of text, which provided enough contexts to support the correctness of the exact answer. Table 2 summarizes all the novelties that have been introduced in the main task over the years of QA campaigns. Each year the tasks proposed were more and more challenging by addressing different types of questions and requiring different types of answer format as output [12]. The principal difference between the QA tracks of CLEF and TREC is the multilingual component of each of its tasks. The combination of the main languages spoken in Europe in the IR has allowed for the creation of multilingual and crosslingual QA systems.

Table 2. Tasks in CLEF QA Track

Task	Characteristics	No. Times performed
Main task	Give the most up-to-date answer supported by the document collection	9
ICLEF (Interactive CLEF)	How best to assist users when searching information written in unknown languages, rather than how best an algorithm can find information written in languages different from the query language	1
QA4MRE (QA for Machine Reading Evaluation)	Focus on the reading of single documents and the identification of the answers to a set of questions about information that is stated or implied in the text	3
WIQA (QA using Wikipedia)	Given a source page from Wikipedia, identify snippets from other Wikipedia pages, possibly in languages different from the language of the source page, that add new and important information to the source page, and without repetition	2
AVE (Answer Validation Exercise)	Decide for given a question and an answer from a QA system, whether the answer is correct or not and it was defined as a problem of recognizing textual entailment in order to promote a deeper analysis in QA	4
QAST (Question Answering on Speech Transcriptions)	Evaluate the task of QA in Speech Transcripts. Accessing information in spoken documents provides additional challenges to those of text-based QA, needing to address the characteristics of spoken language, as well as errors in the case of automatic transcriptions of spontaneous speech.	3
WSD (Word Sense Disambiguation)	Explore the contribution of Word Sense Disambiguation to QA providing document collections and topics which have been automatically tagged with Word Senses from WordNet.	1
RespubliQA	Evaluation task over European legislation	2
Others		5

2.3 NTCIR Conference

Since 2001, NTCIR workshop, co-sponsored by Japan Society for Promotion of Science, has performed specific tracks for the development and evaluation of QA systems. As with CLEF, NTCIR has permitted the evaluation of systems of QA multilingual and crosslingual systems, with the principal difference that the languages used were predominantly found in Asia. Table 3 shows the tasks performed from 2001 to 2010. Although in recent years it had been decided not to have a specific track for QA, in the last NTCIR workshop, NTCIR-11, which took place in December 2014, they resumed evaluation of QA systems with a pilot task called “QA Lab”, focusing on module-based platforms for system performance evaluations and comparisons for solving real-world university entrance exam questions.

Table 3. Tasks in NTCIR

Name of task	Characteristics	No. Times performed
Main task: 3 parts	Give the most up-to-date answer supported by the document collection	2
CLQA (Cross-lingual QA)	Promote research on cross-lingual QA technology mainly for East Asian languages	2
QAC (QA Challenge)	Obtain appropriate answers to given domain independent questions written in NL from a large corpus	4
IR4QA (Information Retrieval for QA)	Evaluate traditional ranked retrieval of documents using well-studied metrics	2
CCLQA (Complex Cross-Lingual QA)	Evaluate QAS on complex	2
QA-Lab Task	Provide a module-based platform for advanced QAS and comparative evaluation for solving real-world university entrance exam questions	1

The evolution of this type of QA track in each competition has been different, CLEF being the conference which has for longest performed specific QA tracks for the evaluation of QA systems. The different QA tracks lead to the opportunity to experiment how the systems and technologies of QA systems respond in different scenarios.

3 Test Questions

With regard to the number of the test questions that have had to be used by participants in the three campaigns herein analyzed so as to evaluate their QA systems, significant differences are seen. While at TREC the number of evaluation questions has varied significantly from one session to another, at both CLEF and NTCIR a stable number has been maintained (200 in the former, and 100 or 300 in NTCIR). Furthermore, in TREC, possibly due to being the first to evaluate QA systems, it is the case that not all the questions proposed to the participants are finally accepted as valid. Indeed, at TREC, at its earliest conferences, questions were rejected for not offering answers with the necessary requirements or for not being within the proceedings.

Given the multilingual and crosslingual tracks of CLEF and NTCIR, the test questions of these conferences have been translated into all the languages involved. Table 4 gives the average number of questions used on each occasion of each of these conferences for the evaluation of QA systems on all the occasions of each of these conferences.

Table 4. Average of test questions per conference

Conference	**Average of questions**
TREC	481
CLEF	300
NTCIR	224

The evaluation campaigns use diverse types of questions (factoid, definition, list, causal, yes/no, among others) which are best adapted to the peculiarities of the tasks of each occasion of the conference. Figure 1 shows the questions proposed by the organizers of the three conferences according to the name they use, ordered by frequency.

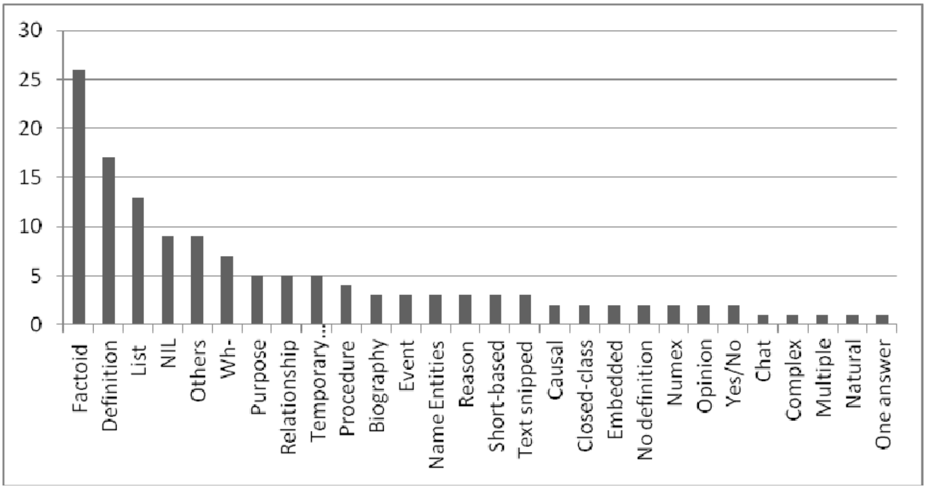

Fig. 1. Frequency each type of test questions is been used

As can be seen, the most used type of question is the factoid question (that is, fact-based questions, asking for the name of a person, a location, the extent of something, the day on which something happened, etc.), followed by definition questions and list questions, respectively. In addition, CLEF is the conference which has applied a greater variety of types of test question (see Table 5), although, as has been stated, it has also been the international forum which has organized the most tracks on the evaluation of QA systems.

Table 5. Type of test questions per conference

Conference	Total of type of questions
CLEF	18
NTCIR	15
TREC	7

4 Answer Assessment

Organizers of the various evaluation campaigns are sent by participants the results given by the latter's systems to the questions put. While at NTCIR the parameters established for the evaluation of the answers have been changed, at the other two conferences the same criteria for all the tasks have been maintained, save for a few exceptions. All answers were judged by native language human assessors, who assigned to each response a unique judgment following a schema pre-established.

In TREC the process was begun with a binary evaluation of the questions (0-correct and 1-incorrect) in order to apply thereafter the already traditional evaluation of correct/right, incorrect/wrong, not exact or not support. The latest QA tasks saw the introduction of the dichotomy "globally correct" or "locally correct" for the correct questions. CLEF has followed the same model as TREC, also used by NTCIR on a few occasions, with the following classification: right (R), wrong (W), unsupport (U) or inexact (X), except in AVE tasks (Answer Validation Exercise) where the answers have been evaluated as validated, unknown or rejected.

5 Evaluation Measures

Several evaluation measures have been used in the QA tasks. In each competition a main measure was selected to rank the results of the participating systems, while several additional measures were adopted in order to provide more information about the systems' performances [2].

After having identified the different evaluation measures proposed in the tasks to be used in the evaluation of the QA systems, it can be seen that around 50 evaluation measures of very different types have been used (see Table 6).

Table 6. Measures used by the conferences

	CLEF	TREC	NTCIR	TOTAL (not repeated)
No. Measures	28	18	18	48

As seen in Table 7, it is the traditional measures of IR (precision, recall, accuracy) which continue to be the most used. However, we find variances in the different tracks since the measures are applied so that they are suited best to the particularities of each task and the evaluation needs of each competition. Another of the commonly used measures is MRR (Mean Reciprocal Rank) proposed by TREC, which is very useful for the evaluation of QA systems as it makes it possible to take into account all the answers retrieved by the system and to assign it a reciprocal value in accordance with the ranking of the system. The traditional F-measure, which combines with recall and precision, is also prominent.

Table 7. Frequency each evaluation measure used

Measure		Frequency of Measure Uses	Frequency by Conferences		
			TREC	CLEF	NTSIR
Accuracy		**26**	**7**	**17**	**2**
	Overall Accuracy	6	-	6	-
	QA rejected accuracy	1	-	1	-
	Random QA Accuracy	1	-	1	-
	QA Accuracy Max	1	-	1	-
Precision		**22**	**7**	**10**	**5**
	Overall Precision	2	-	2	-
	Instance Precision	5	5	-	-
	MAP (Mean Average Precision)	1	1	-	-
	R-Precision	1	1	-	-
	Average Precision	3	1	-	2
Recall		**21**	**8**	**9**	**4**
	Overall Recall	1	-	1	-
	Instance Recall	5	5	-	-
MRR (Mean Reciprocal Rank)		**15**	**3**	**8**	**4**
	RR (Reciprocal Rank)	2	-	-	2
F		**13**	**6**	**7**	**-**
	Overall F	1	-	1	-
	Average F-measure (AFM)	2	-	-	2
	Modified F measure(MF1)	2	-	-	2
	Mean F	6	-	-	6
Confidence Weighted Score (CWS)		9	4	5	-
Final Score		7	7	-	-
Nugget pyramid		7	4	-	3
C@1		5	-	5	-
Ave Length		5	5	-	-
K1		4	-	4	-
Exact match		3	-	-	3
Softmatch		3	-	-	3
aBinarized		3	-	-	3
Nugget recall		3	-	-	3
Q		2	-	-	2
Top 5		2	-	-	2
Ndcg (Normalized Discounted Cumulative Gain)		2	-	-	2
Error rate		2	2	-	-
BA-HIT@1		1	-	-	1
P@N		1	-	1	-
Yield		1	-	1	-
Px		1	-	-	1
ES (Expression set)		1	-	-	1
CAS (Correct answer set)		1	-	-	1
UR /t		1	-	-	1
MRC (Mean Reciprocal Cost)		1	-	-	1
Px		1	-	-	1
FHS (First Hit Success)		1	1	-	-

Table 7. (*continued*)

Qrels	1	1	-	-
CR (Correct Test)	1	-	-	1
PMM (Population Marginal Mean)	1	1	-	-

As stated, each conference has its own objectives, tradition and history. This entails that not all the evaluation measures are equally applicable to the distinct tracks of the different campaigns analysed. However, space could have been found for some of those measures at several of those conferences. That is the case with Instance Precision, Mean Average Precision and R-Precision, which have only been applied in TREC. There is a similar situation with Error Rate, with the measure which determines the average length of the answers (Ave Length) and con FHS (First Hit Succes), amongst others.

On the other hand, possibly due to the fact that CLEF is a campaign with more delimited objectives, that is to say, essentially oriented at the evaluation of systems types, the multilingual and the crosslingual, this conference introduces some specific measures albeit some of them with notably reduced frequency. For example, in relation to the measure Accuracy, others such as Overall Accuracy, QA rejected accuracy, Random QA Accuracy and QA Accuracy Max have been proposed.

However, NTSIR is the conference with the greater number of specific measures (exact match, softmatch, aBinarized, Normalized Discounted Cumulative Gain, Correct answer set, amongst others) implemented in the evaluations of its tracks, and while some are found in up to six campaigns, the majority have been applied infrequently.

6 Conclusions

Evaluation of Information retrieval evaluation is a constantly evolving field [14-15]. In this study we have set out the QA tasks or specific labs organized in the three principal IR evaluation campaigns over the last 15 years. Likewise, the type of test question used and the main evaluation measures proposed have been identified.

The new scenario raised by the appearance of novel QA systems constitutes an excellent benchmaking to test the relevance of the traditional evaluation and to propose and to adapt new methods and measures appropriate for this environment.

References

1. Dwivedi, S.K., Singh, V.: Research and Reviews in Question Answering System. Procedia Technology 10, 417–424 (2013)
2. Peñas, A., Magnini, B., Forner, P., Sutcliffe, R., Rodrigo, A., Giampiccolo, D.: Question answering at the cross-language evaluation forum 2003—2010. Lang. Resour. Eval. 46(2), 177–217 (2012)

3. Mollá, D., Vicedo, J.L.: Question Answering in Restricted Domains: An Overview. Comput. Linguist. 33(1), 41–61 (2007)
4. Diekerma, A.R., Yilmazel, O., Liddy, E.D.: Evaluation of restricted domain Question-Answering systems Center for Natural Language Processing. Paper 3 (2004)
5. Sing, G.O., Ardil, C., Wong, W., Sahib, S.: Response Quality Evaluation in Heterogeneous Question Answering System: A Black-box Approach. In: Proceedings of World Academy of Science, Lisbon, vol. 9 (2005)
6. Olvera-Lobo, M.D., Gutiérrez-Artacho, J.: Question-answering systems as efficient sources of terminological information: an evaluation. Health Information & Libraries Journal 27, 268–276 (2010)
7. Olvera-Lobo, M.D., Gutiérrez-Artacho, J.: Language resources used in multi‑lingual question‑answering systems. Online Information Review 35(4), 543–557 (2011)
8. Olvera-Lobo, M.D., Gutiérrez-Artacho, J.: Multilingual Question-Answering System in biomedical domain on the Web: an evaluation. In: Forner, P., Gonzalo, J., Kekäläinen, J., Lalmas, M., de Rijke, M. (eds.) CLEF 2011. LNCS, vol. 6941, pp. 83–88. Springer, Heidelberg (2011)
9. Olvera-Lobo, M.D., Gutiérrez-Artacho, J.: Evaluation of Open- vs. Restricted- Domain Question Answering Systems in the Biomedical Field. Journal of Information Science 37(2), 152–162 (2011)
10. Olvera-Lobo, M.D., Gutiérrez-Artacho, J.: Evaluación del rendimiento de los sistemas de búsqueda de respuestas de dominio general. Revista española de Documentación Científica 36(2), e9 (2013)
11. Radev, D.R., Qi, H., Wu, H., Weiguo, F.: Evaluating Web-based Question Answering Systems. In: Proceedings LREC (2002)
12. Forner, P., Giampiccolo, D., Magnini, B., Peñas, A., Rodrigo, Á., Sutcliffe, R.: Evaluating multilingual question answering systems at CLEF. In: Proc. 7th International Conference on Language Resources and Evaluation. LREC 2010, Malta, pp. 2774–2781 (2010)
13. Voorhees, E.: Overview of TREC 2003. In: Proceedings of the 12th Text Retrieval Conference (2003)
14. Kelly, D., Sugimoto, C.R.: A systematic review of interactive information retrieval evaluation studies, 1967–2006. Journal of the American Society for Information Science and Technology 64, 745–770 (2013)
15. Moghadasi, S.I., Ravana, S.D.: Low-cost evaluation techniques for information retrieval systems: A review. Journal of Informetrics 7(2), 301–312 (2013)