New Measures for Open-Domain Question Answering Evaluation within a Time Constraint

Elisa Noguera[1], Fernando Llopis[1], Antonio Ferrández[1], and Alberto Escapa[2]

[1]GPLSI. Departamento de Lenguajes y Sistemas Informáticos
Escuela Politécnica Superior. University of Alicante
{elisa,llopis,antonio}@dlsi.ua.es
[2]Departamento de Matemática Aplicada
Escuela Politécnica Superior. University of Alicante
alberto.escapa@ua.es

Abstract. Previous works on evaluating the performance of Question Answering (QA) systems are focused on the evaluation of the precision. In this paper, we developed a mathematic procedure in order to explore new evaluation measures in QA systems considering the answer time. Also, we carried out an exercise for the evaluation of QA systems within a time constraint in the CLEF-2006 campaign, using the proposed measures. The main conclusion is that the evaluation of QA systems in realtime can be a new scenario for the evaluation of QA systems.

1 Introduction

The goal of Question Answering (QA) systems is to locate concrete answers to questions in collections of text. These systems are very useful for the users because they do not need to read all the document or fragment to obtain a specific information. Questions as: How old is Nelson Mandela? Who is the president of the United States? When was the Second World War? can be answered by these systems. They contrast with the more conventional Information Retrieval (IR) systems, because they treat to retrieve relevant documents to a query, where the query may be a simply collection of keywords (e.g. old Nelson Mandela, president United States, Second World War, ..).

The annual Text REtrieval Conference (TREC[1]), organized by the National Institute of Standards and Technology (NIST), is a serie of workshops designed to advance in the state-of-the-art in text retrieval by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. This model has been used by Cross-Language Evaluation Forum (CLEF[2]) in Europe and by the National Institute of Informatics Test Collection for IR Systems (NTCIR[3]) in Asia, which have also studied the cross-language issue. Since 1999, TREC have a specific QA track ([3]). CLEF ([2]) and NTCIR ([1]) have also introduced the

[1] http://trec.nist.gov
[2] http://www.clef-campaign.org
[3] http://research.nii.ac.jp/ntcir

QA evaluation. This evaluation consists of given a large number of newspaper and newswire articles, participating systems try to answer a set of questions by analyzing the documents in the collection in a fully automated way.

The main evaluation measures used in these forums are *accuracy* evaluation measure, but different metrics were also considered: *Mean Reciprocal Rank (MRR)*, *K1 measure* and *Confident Weighted Score (CWS)* (for further information about these metrics see [2]).

The motivation of this work is to study the evaluation of QA systems within a time constraint. In order to evaluate the answer time of the systems and compare them, we carried out an experiment in the CLEF-2006 providing a new scenario in order to compare the QA systems. Specifically, we have proposed new measures to evaluate not only the effectiveness of the systems, but also the answer time. As the results achieved by the systems, we can argue that this is a promising step to change the direction of the evaluation in QA systems.

The remainder of this paper is organized as follows: next section presents presents a new proposal of evaluation measures for QA systems. Section 3 describes the experiment carried out in the QA context at CLEF-2006, the evaluation used and the results achieved. Finally, section 4 gives some conclusions and future work.

2 New approaches evaluating QA systems

The above mentioned problem can be reformulated in a mathematical way. Let us consider that the answer of each system S_i can be characterized for our purposes by an ordered pair of real numbers (x_i, t_i). The first element of the pair reflects the precision of the answer and the second one the efficiency. In this way, a QA task can be represented geometrically as a set of points located in a subset $D \subseteq \mathbb{R}^2$. Our problem can be solved by giving a method that allows to rank the systems S_i accordingly to some prefixed criterion that take into account both the precision and the efficiency of each answer. This problem is of the same nature as others tackled in Decision theory.

A solution to this problem can be achieved by introducing a total preorder, sometimes referred as total quasiorder, in D. Let us remind you that a binary relation $\preceq$ on a set D is a total preorder if it is reflexive, transitive and any two elements of D are comparable. In particular, we can define a total quasiorder on D with the aid of an auxiliary two variables real function $f : D \subseteq \mathbb{R}^2 \rightarrow I \subseteq \mathbb{R}$, in such a way that:

$$(a, b) \preceq (c, d) \quad \Leftrightarrow f(a, b) \leq f(c, b), \, \forall \, (a, b), \, (c, d) \in D. \tag{1}$$

We will refer to this function as a ranking function. One of the advantages of this procedure is that the ranking function contains all the information relative to the chosen criterion to classify the different systems S_i. Anyway, let us underline that, since the binary relation $\preceq$ defined in this way is not necessarily an order in D, two different elements of D can be equal with respect to the preorder $\preceq$, that is, they are in the same position of the ranking. Mathematically all the elements that are tied in the classification belong to a level curve of the ranking function, that we will call iso-ranking curve. Namely, the iso-ranking curves are characterized by all the elements of D that fulfill the equation $f(x,t) = L$, being L a real number in the image of f, I. Let us point out that the proposed ranking procedure to evaluate the QA task is of an ordinal type. This means that we should not draw a direct conclusion about the absolute difference of the numerical values of the ranking function for two systems. The only relevant information concerns to their relative position in the relative ranking of the QA task. As a matter of fact, if we consider a new ranking function constructed by composing the initial ranking function with a strictly increasing function, the numerical value assigned to each system is changed but the final ranking obtained is the same as the first one.

In the approach developed in this paper, the precision of the system S_i is given the mean reciprocal rank (MRR), so $x_i \in [0,1]$. The efficiency is measured by considering the answer time of each system, in such a way that a smaller time to answer means a better efficiency of a system. Anyway, to obtain a more suitable scale of representation, we have considered the effective time resulting from dividing the answer time by the maximum answer time obtained in the QA task under consideration, hence we will have that this effective time, denoted as t_i, belong to the interval $(0,1]$. In this way, the accessible region of $\mathbb{R}^2$ is given by the set $D \equiv [0,1] \times (0,1]$.

To define a realistic ranking function, it is necessary to require to this function some additional features. These properties are based on the intuitive behavior that our ranking criterion should have to fulfill. For example, as a preliminary approach, we are going to demand the ranking function that:

1. The function f must be continuous in D.
2. The supremum of I is given by $\lim_{t \to 0} f(1,t)$. In the case that I is not upper bound, we must have $\lim_{t \to 0} f(1,t) = +\infty$.
3. The infimum of I is given by $f(0,1)$.

The first condition is imposed for mathematical convenience, although it can be interpreted in terms of some simplified arguments. Namely, this requirement excludes the possibility that if two systems are in different positions in the ranking, any arbitrarily small variation in the precision or the efficiency of one of them changes their relative positions. The second condition is related with the fact that the fictitious system defined by the pair $(1,0)$ always must be in the first position of the ranking. Finally, the last condition implies that the pair $(0,1)$ must be in the last position.

2.1 Ranking function independent of time (MRR_2)

As a first simple example of ranking function, let us consider $MRR_2(x,t) = x$. The preorder induced by this function is closed to the lexicographical order, some times called alphabetic order. For this ranking function we have that:

1. The image of MRR_2 is the interval $[0,1]$.
2. The function MRR_2 is continuous in D.
3. $\lim_{t\to 0} MRR_2(1,t) = 1$.
4. $MRR_2(0,1) = 0$.

So, this function fulfills all the previous requirements. On the other hand, the iso-ranking curves of the function are of the form $x = L$, $L \in [0,1]$ whose representation is a family of vertical segments of length unity (see figure 1). The preorder constructed from this ranking function only takes into account the precision, being unaware of the efficiency of the systems.

2.2 Ranking function with inverse temporal dependence ($MRRT$)

As a second example of ranking function that does take into account the efficiency of the systems, we are going to consider $MRRT(x,t) = x/t$. Let us note that in this case the ranking function is inversely proportional to the time and directly proportional to the precision. In particular, this function verifies the properties:

1. The image of $MRRT$ is the interval $[0,+\infty)$.
2. The function $MRRT$ is continuous in D.
3. $\lim_{t\to 0} MRRT(1,t) = +\infty$.
4. $MRRT(0,1) = 0$.

The associated iso-ranking curves to the function are of the form $x/t = L$, $L \in [0,+\infty)$. Geometrically these curves are a family of segments passing through the point $(0,0)$ and with slope $1/L$ (see figure 1). In this way, the systems with better efficiency, that is, smaller effective time, obtain for a given value of x a large value of the ranking function. As a matter of fact, both precision and efficiency have the same influence on the ranking function, since a system of values (x,t) is tied with the system $(\alpha x,\alpha t)$ with $0 < \alpha < 1$. On the other hand, although the information of the ranking function is of an ordinal nature, it is desirable that the image of the function is between 0 and 1, since this facilitates an intuitive representation of the values of the ranking function, a condition that this function does not verify either.

2.3 Ranking function with inverse exponential-like with time dependence $MRRT_e$

Due to the disadvantages of the previous functions, we propose a new ranking function that depends both on the precision and the efficiency of the system but in which the efficiency has less weight than the precision when evaluating QA systems. Namely, we are going to introduce the ranking function

$$MRRT_e(x,t) = \frac{2x}{1+e^t}, \tag{2}$$

being e^t the exponential of the effective time. This function fulfills the following requirements:

1. The image of $MRRT_e$ is the interval $[0, 1)$.
2. The function $MRRT_e$ is continuous in D.
3. $\lim_{t \to 0} MRRT_e(1, t) = 1$.
4. $MRRT_e(0, 1) = 0$.

The iso-ranking curves of this function are of the form $2x/(1 + e^t) = L$, $L \in [0, 1)$, whose representation is sketched in figure 2. Let us underline that for fictitious efficient systems, that is, those systems that answer instantaneously ($t = 0$), this ranking function coincides with the usual precision classification. Nevertheless, the functional dependence on time modulates the value of x, in such a way that when the time grows up the value of the ranking function, for a fixed precision, decreases. Anyway, this modulation is smoother than in the case of the ranking function inversely proportional to time. Moreover, if we consider a given system S, we can only tie with it by considering systems whose precision, and efficiency, vary on a particular range, not for any arbitrarily small value of the precision.

3 Experiment at QA CLEF-2006

As above is mentioned, we considered the time as a fundamental part in the evaluation of QA systems. In accordance with CLEF organization, we carried out a pilot task at CLEF-2006 whose aim was to evaluate the ability of QA systems to answer within a time constraint. This is an innovative experiment and the initiative is aimed towards providing a new scenario for the evaluation of QA systems. This experiment follows the same procedure that the QA@CLEF-2006, but the main difference is the consideration of the answer time.

In total, five groups took part in this pilot task. The participating groups were: *daedalus* (Spain), *tokyo* (Japan), *priberam* (Portugal), *alicante* (Spain) and *inaoe* (Mexico). All of them participated in the main QA task at CLEF-2006, and have experience researching in QA systems.

3.1 Performance Evaluation

In this section we present the evaluation results of the five systems which participated in the realtime experiment. On the one hand, we present precision and the efficiency obtained by these systems. On the other hand, we present the score achieved by them with the different metrics which combine the precision with the efficiency. Tables 1 shows the summary of results for the used metrics (MRR, t, MRRT, $MRRT_e$).

The precision of the QA systems was evaluated in the experiment with the MRR metric. It is presented in table 1. As above is mentioned, the efficiency of the systems is measured with the answer time (in seconds). In order to normalize the obtained answer times between 0 and 1, we use t as $tsec/tmax$ ($tsec$ is the answer time in seconds and $tmax$ is the highest answer time value of the list).

Table 1. Evaluation results with the different metrics

Participant	MRR	rank	t	rank	MRRT	rank	$MRRT_e$	rank
daedalus1	**0.41**	**1°**	0.10	4°	3.83	4°	**0.3881**	**1°**
tokyo	0.38	2°	1.00	6°	0.38	6°	0.2044	6°
priberam	0.35	3°	**0.01**	**1°**	**32.13**	**1°**	0.3481	**2°**
daedalus2	0.33	4°	0.03	3°	8.56	3°	0.3236	3°
inaoe	0.3	5°	0.38	5°	0.78	5°	0.2433	**4°**
alicante	0.24	6°	0.02	2°	16.23	2°	0.2382	**5°**

The overall evaluation of the QA systems, combining precision with the answer time with the MRR_2 metric (see section 2) is the same than the evaluation of the MRR (see section 1), because this measure takes into account the precision firstly, and it takes only into account the time if the precision among two or more systems is the same. Graphically, an iso-ranking curve is made up of all the systems with the same value of x and arbitrarily different values of the effective time. That is, the ranking criterion is the same as the usual performance followed up today to evaluate QA systems. The limitations of this procedure, which have motivated this work, are clear if we consider the systems tokio and priberam in figure 1. With the above considered ranking function the system tokyo is in the second position of the ranking and the system priberam in the third one. However, the difference in the precision (MRR) of the systems is very small, 0.38 vs. 0.35 , whereas the efficiency of the system priberam is much better than the efficiency of the system tokyo. Therefore, it would be reasonable that the system priberam was preferred to the system tokyo. This is impossible with this kind of ranking functions, since they are independent of time.

The evaluation of the systems with the MRRT metric (see section 2) is presented in table 1. It shows by each system, the rank in the list that it has obtained with the MRRT measure. Graphically, we can see the different obtained values in the figure 1. For example, the system alicante whose precision is 0.24 of x and 0.02 of t is in the same position of the ranking as priberam whose precision is better (0.35). The position of any system in the ranking can be always tied with a system of smaller precision but larger efficiency, in particular this can be taking any arbitrarily small precision value. This is a disadvantage because the efficiency of the systems are taken too much into account and, in our opinion, the leading factor to evaluate QA systems must be the precision of the answer.

The evaluation of the systems with the $MRRT_e$ measure is also presented in table 1. daedalus1 and priberam obtain the best results of $MRRT_e$ (0.3881 and 0.3481 respectively). The decrease in MRR of priberam (from 0.35 to 0.3481) is not significant because it has a short answer time (76 seconds), just like alicante (from 0.24 to 0.2382). Nevertheless, the $MRRT_e$ of daedalus1 reduces the MRR (from 0.41 to 0.3881), because it has a upper answer time (549 seconds). Finally, inaoe and tokyo are significantly penalized because their t are higher. Graphically, we can compare the different values of $MRRT_e$ in the figure 2, seeing that the fastest systems (priberam and alicante) have a similar performace than evaluating only the MRR. However, inaoe and tokyo are penalized obtaining lower results than evaluating only the MRR. For example, tokyo had the second best MRR (0.38) and the worst t, and it is penalized being the last in the ranking of $MRRT_e$. Also, to obtain the same position in the ranking as the system $S \equiv (0.4, 0.2)$ the precision needed could vary in the range from 0.36 to 0.67, corresponding to a variation of the time from 0 to 1. These particularities makes the ranking function $MRRT_e$ very suitable for the evaluation of QA systems in realtime.

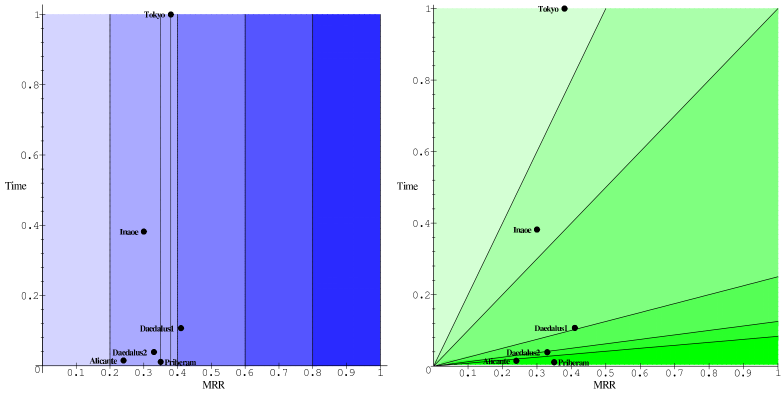

Fig. 1. Comparatives of the results obtained by each system with the Lexicographical preorder and $MRRT$ evaluation measures respectively in its iso-ranking function.

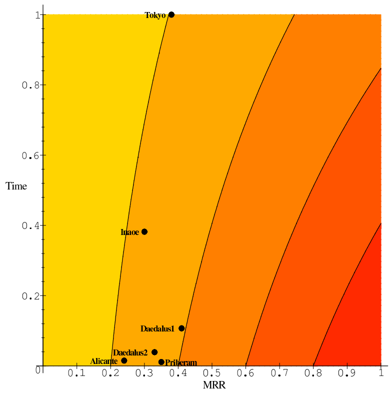

Fig. 2. Comparative of the results obtained by each system with the $MRRT_e$ evaluation measure in its iso-ranking function.

4 Conclusions and Future Work

Mainly, the evalution of QA systems is studied deeply in three known evaluation forums: TREC, CLEF and NTCIR. But, these forums are only focused on evaluating the precision of the systems, and they do not evaluate their efficiency (we consider the answer time of the system as measure of efficiency). Mostly, this evaluation entails accurate systems but slowly at the same time. For this reason, we studied the evaluation of QA systems taking into account the answer time.

For the evaluation of the QA systems, we proposed three measures (MRR_2, MRRT, $MRRT_e$) to evaluate them within a time constraint. These measures are based on the Mean Reciprocal Rank (MRR) and the answer time. As preliminary results, we show that $MRRT_2$ only takes into account the precision and the measure MRRT takes into account too much the time. We have solved this inconvenience proposing a new measure called $MRRT_e$. It also combines the MRR with the answer time, but it is based on an exponential function. It penalizes the systems that has a higher answer time. In conclusion, the new measure MRR_e allows classify the systems considering the precision and the answer time.

Futhermore, we carried out a task in the CLEF-2006 in order to evaluate QA systems within a time constraint. This is the first evaluation of QA systems in realtime. It has allowed to stablish a methodology and criterion for the evaluation of QA systems in a new scenario. Fortunately, this exercise did receive a great attention by both organizers and participants, because of seventeen groups were interested into participating and the exercise and the presentation of the results in the workshop were very successful.

Finally, the future directions that we plan to undertake are to take into account more variables as the hardware used by the systems, as well as to insert new control parameters in order to give more significance to efficiency or precision.

5 Acknowledgments

This research has been partially supported by the framework of the project QALL-ME (FP6-IST-033860), which is a 6th Framenwork Research Programme of the European Union (EU), by the Spanish Government, project TEXT-MESS (TIN-2006-15265-C06-01) and by the Valencia Government under project number GV06-161.

References

1. J. Fukumoto and et al. Question Answering Challenge (QAC-1). In *Proceedings of the Third NTCIR Workshop*, 2002.
2. B. Magnini and et al. Overview of the CLEF 2006 Multilingual Question Answering Track. In *WORKING NOTES CLEF 2006 Workshop*, 2006.
3. E. M. Voorhees and H. T. Dang. Overview of the TREC 2005 Question Answering Track. In *TREC*, 2005.