How long can you wait for your QA system?

Fernando Llopis
Departamento de Lenguajes y Sistemas Informáticos
Escuela Politécnica Superior
University of Alicante, Spain
llopis@dlsi.ua.es

Alberto Escapa
Departamento de Matemática Aplicada
Escuela Politécnica Superior
University of Alicante, Spain
alberto.escapa@ua.es

Antonio Ferrández
Departamento de Lenguajes y Sistemas Informáticos
Escuela Politécnica Superior
University of Alicante, Spain
antonio@dlsi.ua.es

Sergio Navarro
Departamento de Lenguajes y Sistemas Informáticos
Escuela Politécnica Superior
University of Alicante, Spain
snavarro@dlsi.ua.es

Elisa Noguera
Departamento de Lenguajes y Sistemas Informáticos
Escuela Politécnica Superior
University of Alicante, Spain
elisa@dlsi.ua.es

ABSTRACT

Common approaches to evaluate Question Answering (QA) systems consider exclusively the accuracy of the answers. It ignores an essential feature of all the computational procedures: the efficiency. In this note, we explore new evaluation measures that take into account, in addition to the accuracy, the efficiency, which is incorporated through the magnitude of the answer time of QA systems. In particular, we have developed a family of metrics where the signification of the efficiency can be balanced. By applying this metric to a real time experiment performed in CLEF 2006, it is showed different possibilities to evaluate in a more realistic way the performance of QA systems.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

Keywords

Question Answering, Performance, Evaluation Measures

1. INTRODUCTION

The main evaluation measures used for QA systems are *accuracy*, or some related metrics such as *Mean Reciprocal Rank (MRR)*, *K1 measure* and *Confident Weighted Score (CWS)*. In any case, the answer time of the QA systems is not considered, that is to say, it is neglected the efficiency of the systems. By so doing, we face two main difficulties: some systems can have a good performance being extremely slow in obtaining the right answers; and the comparison among QA systems is not realistic when they had employed different answer times. Therefore, a realist performance analysis requires to take into account the accuracy of the answers and the time needed to obtain them. The aim of this note is to develop a metric that considers these two properties of QA systems, in such a way that the user can balance the dependence of the metric on the efficiency of the system.

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.

2. NEW EVALUATION MEASURES BASED ON ANSWER TIME

One simple possibility to define a metric depending on the accuracy and the efficiency of a system is to associate two real numbers, x and t, to each of these characteristics. Then, we can construct a real function f of two independent real variables and order the systems accordingly the values obtained when evaluating $f(x,t)$. We refer to f as a ranking function, since it allows ranking the different systems depending on their accuracy and answer time. This approach also provides a graphical view of the ordering procedure of the systems through the level curves of f, which we will call iso-ranking curves. Mathematically all the systems that are tied in the classification belong to the same level curve. In the case of accuracy based metrics the level curves are vertical straight lines increasing from left to right, but when the metric also considers the efficiency this is not longer true. We can view an example for one particular metric $MRRT_{E,1}$ (see the next section) in figure 1.

It is important to note that this procedure is of an ordinal type. This means that the relevant information to classify the systems is the relative difference of the numerical values of the ranking function, being meaningless the concrete value of the ranking function for a single system. On the other hand, the ranking functions are not completely arbitrarily but must fulfill some mathematical requirements ([1]).

Within this framework there have been considered different kinds of ranking functions ([1]), in such a way that the efficiency has less weight than the accuracy, since by no means a completely inaccuracy system is preferred over a very efficient one. Anyway, it is possible to modulate the weight of the efficiency in the evaluation of QA systems. To this end, we have introduced a family of ranking functions of the same type controlled by a parameter. By so doing, the value of the parameter could be adjusted in any QA task allowing to design different evaluation measures, accordingly some prefixed criteria. In particular, we have constructed a

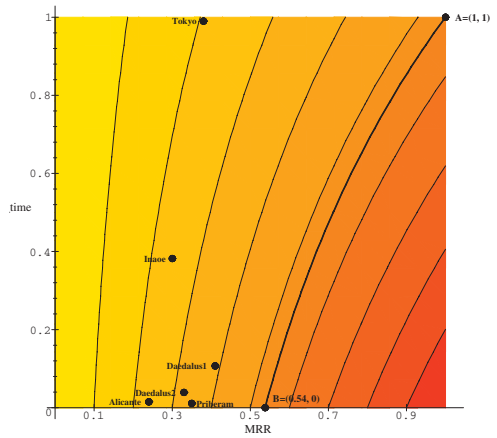

Figure 1: Iso-ranking curves for CLEF-2006 results with metric $MRRT_{E,1}$.

family of ranking functions of the form

$$MRRT_{E,r}(x,t) = \frac{2x}{1+e^{rt}}. \quad (1)$$

Here, the accuracy of the system x is given the mean reciprocal rank (MRR), so $x \in [0,1]$. 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, belongs to the interval $(0,1]$. Finally, r denotes the parameter that controls the efficiency dependence.

If we take $r = 0$ we recover the MRR measure, which only takes into account the accuracy of the system. In general, the real parameter r can only take values in the interval $[0,+\infty)$. When the value of r increases from 0 to $+\infty$ the weight of the efficiency is also increased. In this way, a ranking function with a small value of the parameter r takes into account very little the efficiency of the systems. This is clear if we observe the functional form of the ranking function family, where the MRR value is multiplied by a function that only depends on time and always take positive values equal or smaller than 1. For higher values of r the value of MRR is more and more penalized as the time grows up.

3. DISCUSSION

Next, we analyze an application of the above designed metric to a real evaluation scenery. 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, in others words, to consider the efficiency as a relevant part in the evaluation. This experiment followed the same procedure that the main task at QA@CLEF-2006, but the main difference was the consideration of the answer time. The participating groups were: *daedalus* (Spain), *tokyo* (Japan), *priberam* (Portugal), *alicante* (Spain) and *inaoe* (Mexico) (for further information about the realtime experiment see [2]). In table 1, the results of the competition are displayed. We have evaluated

Table 1: CLEF-2006 results

Team	MRR	t (s)	Ef. time
daedalus1	0.41	549	0.10
tokyo	0.38	5141	1.00
priberam	0.35	56	0.01
daedalus2	0.33	198	0.03
inaoe	0.3	1966	0.38
alicante	0.24	76	0.02

the performance of these teams with the uniparametric family of evaluation measures $MRRT_{E,r}$. In this way, it is possible to obtain different classification of the systems determined by the values of the parameter r (see table 2). For example, daedalus1 and tokyo obtain the best results of $MRR = MRRT_{E,0}$ (0.41 and 0.38 respectively). But, the position of tokyo goes down in the ranking accordingly we increase the values of r, that is to say, when the answer time becomes more important. On the contrary, alicante obtains the worst value of MRR (0.24), as a consequence it is the last one in the ranking if we take only the MRR into account, but it goes up if we increase the parameter r. The teams daedalus1 and priberam do not change practically their position in the ranking, although if we increase the parameter r their values bring near, because priberam has a shorter answer time than daedalus1.

Table 2: Accuracy-efficiency evaluation

Participant	r=0	r=0.51	r=0.99	r=1.95
daedalus1	**0.41 (1°)**	0.40 (1°)	0.39 (1°)	0.37 (1°)
tokyo	0.38 (2°)	**0.28 (4°)**	**0.19 (6°)**	0.09 (6°)
priberam	0.35 (3°)	**0.35 (2°)**	0.35 (2°)	0.35 (2°)
daedalus2	0.33 (4°)	**0.33 (3°)**	0.32 (3°)	0.32 (3°)
inaoe	0.30 (5°)	0.27 (5°)	0.23 (5°)	0.19 (5°)
alicante	0.24 (6°)	0.24 (6°)	**0.24 (4°)**	0.24 (4°)

Summarizing up, we have proposed a procedure to define different metrics that consider both the accuracy and efficiency of QA systems and that allows to control the weight of the efficiency on the metric. It opens a new line beyond the traditional evaluation paradigm, since efficiency of QA systems should not be longer ignored.

4. ACKNOWLEDGEMENTS

This paper has been partially supported by the Spanish government, project TIN-2006-15265-C06-01, and by the framework of the project QALL-ME, which is a Sixth Framework Research Programme of the European Union (EU), contract number: FP6-IST-033860.

5. REFERENCES

[1] Noguera, E., Llopis, F., Ferrández, A. and Escapa, A. New Measures for Open-Domain Question Answering Evaluation Within a Time Constraint. Lecture Notes in Computer Science, 4629, 540–547, 2007.

[2] Magnini, B., *et al.*: Overview of the CLEF 2006 Multilingual Question Answering Track. In: WORKING NOTES CLEF 2006 Workshop, 2006.