Question Answering Pilot Task at CLEF 2004

Jesús Herrera, Anselmo Peñas and Felisa Verdejo
Dpto. Lenguajes y Sistemas Informáticos, UNED
{*jesus.herrera, anselmo, felisa*}*@lsi.uned.es*

Abstract

A Pilot Question Answering Task has been activated in the Cross-Language Evaluation Forum 2004 with a twofold objective. In the first place, the evaluation of Question Answering systems when they have to answer conjunctive lists, disjunctive lists and questions with temporal restrictions. In the second place, the evaluation of systems' capability to give an accurate self-scoring about the confidence on their answers. In this way, two measures have been designed to be applied on all these different types of questions and to reward systems that give a confidence score with a high correlation with the human assessments. The forty eight runs submitted to the Question Answering Main Track have been taken as a case of study, confirming that some systems are able to give a very accurate score and showing how the measures reward this fact.

1 Introduction

A Pilot Question Answering (QA) Task has been activated this year within the Main QA Track of the CLEF[1] 2004 competition. The Pilot Task aims at investigating how QA systems are able to cope with another type of questions than the ones posed in the Main Track. To accomplish it, a set of questions has been prepared and new evaluation measures have been proposed.

Few questions were similar to those posed in the Main Track (factoid and definition questions) although they were selected with more than one correct and distinct answer. Questions whose answer is a list of items were also posed, following TREC[2] and NTCIR[3] previous experiences. Finally, more than half of the questions in the Pilot Task aim at dealing with temporal restrictions.

The evaluation measure proposed for this Pilot Task has been designed to take into consideration all these types of questions and, simultaneously, reward systems that, even focusing their attention in a few types of questions, are able to obtain very accurate results, with a good answer validation and a good confidence score.

In the present edition, the Pilot Task has been activated only for Spanish and has been carried out simultaneously with the Main QA Track. Participants in the Pilot Task have made a special effort to accomplish the extra work.

Section 2 describes the task and the different types of questions, including those with temporal restrictions. Section 3 presents some criteria to design the evaluation measure and presents the *K* and *K1* measures. The results for the Main QA Track at CLEF [6] are taken as a case of study to discuss and compare these measures with the previous ones used at TREC, NTCIR and CLEF. Section 4 presents the results obtained by participants in the Pilot Task and, finally, Section 5 points out some conclusions and future work.

[1] Cross-Language Evaluation Forum, http://www.clef-campaign.org
[2] Text REtrieval Conference, http://trec.nist.gov
[3] NII-NACSIS Test Collection for IR Systems, http://research.nii.ac.jp/ntcir/index-en.html

2 Task Definition

The QA Pilot Task followed the rules stated in the QA Main Track guidelines except for the source and the target languages, the type and number of questions, and the evaluation measure.

One hundred of questions were posed in Spanish and the corpus used was the EFE Spanish press agency collection of news from 1994 and 1995. The questions of this Pilot Task were distributed throughout the following types: factoid (18), definition (2), conjunctive list (20), temporally restricted by date (20), temporally restricted by period (20), and temporally restricted by event (20 nested questions). A little amount of questions had no answer in the document collection (2 NIL factoid questions). As usual, a question was assumed to have no answer when neither human assessors nor participating systems could find one.

Ideally, QA systems should tend to give a unique answer for each question but, however, there exist some questions whose answer depends on the context or evolves in time. In these cases, *disjunctive lists* are obtained, that is, lists of different and correct items representing a disjunction of concepts. The decision of which one of them is the most correct is strongly dependant on the user's information need, text errors, consistency between different texts (specially in the news domain), etcetera. Therefore, being able to obtain all the possible correct and distinct answers for a question seems to be a desirable feature for open domain QA systems.

For this reason, there was no limit for the number of answers at the Pilot Task, but one answer for each question must be given at least. If systems believed that it was no response to a question in the corpus, they had to answer NIL.

In the *conjunctive list* type of questions, a determined or undetermined quantity of items is required for conforming an only answer. A conjunctive list is a series of items representing a conjunction of concepts. For the Pilot Task, the goal was to obtain the largest amount of different items within each answer.

Three subtypes of *temporally restricted* questions have been proposed at the Pilot Task, and three moments with regard to the restriction (before, during or after the temporal restriction):

- **Restriction by Date,** where a precise date contextualises the question, which can refer either to a particular moment, before or after. A date could consist in a day, a month, a year, etcetera, depending on the question. Examples:

  `- T ES ES 0011 ¿Qué sistema de gobierno tenía Andorra hasta mayo de 1993?`

  `- T ES ES 0014 ¿Quién visitó Toledo el 22 de febrero de 1994?`

- **Restriction by Period.** In this case, questions are referred explicitly to a whole period or range of time. A period could be expressed by a pair of dates delimiting it, or by a name accepted as designation of some important periods as, for example, *Cuaresma*[4]. Examples:

  `- T ES ES 0086 ¿Quién reinó en España durante el Siglo de Oro de su literatura?`

  `- T ES ES 0037 ¿Quién gobernó en Bolivia entre el 17 de julio de 1980 y el 4 de agosto de 1981?`

- **Event restriction,** that implies an embedded or implicit extra question because it is necessary to answer the nested question to determine the temporal restriction. Then, the temporal restriction refers to the moment in which the second event occurred. For example:

  `- T ES ES 0098 ¿Quién fue el rey de Bélgica inmediatamente antes de la coronación de Alberto II?`

  `- T ES ES 0079 ¿Qué revolución estudiantil surgió en Francia al año siguiente de la Guerra de los Seis Días?`

[4]Cuaresma is the Spanish word standing for Lent.

The degree of inference necessary to solve the temporal restrictions was not the same for all the questions. In some questions a reference to the temporal restriction could be found in the same document, while in other questions it was necessary to accede to other documents to temporally locate the question.

3 Evaluation Measure

The evaluation measure has been designed in order to reward systems that return as many different and correct answers as possible to each question but, at the same time, punishing incorrect answers. Two reasons motivate the negative adding for the incorrect answers: First, it is assumed that a user of a QA system would prefer a void answer rather than an incorrect one. Systems must validate their answers and must give an accurate confidence score. Second, since there was no limit in the number of answers, systems must calibrate the risk of giving too much incorrect ones. The effect was that no more than three answers per question were given.

In order to evaluate systems' self-scoring, a mandatory confidence score given by means of a real number ranged between 0 and 1, was requested. 0 meant that the system had no evidence on the correctness of the answer, and 1 meant that the system was totally sure about its correctness.

The evaluation measure has been designed to reward systems that:

- answer as many questions as possible,
- give as many different right answers to each question as possible,
- give the smaller number of wrong answers to each question,
- assign higher values of the score to right answers,
- assign lower values of the score to wrong answers,
- give answer to questions that have less known answers.

3.1 The K-measure

According to the criteria above, the evaluation measure is defined as follows:

$$K(sys) = \frac{1}{\#questions} \cdot \sum_{i \in questions} \frac{\sum_{r \in answers(sys,i)} score(r) \cdot eval(r)}{max\{R(i), answered(sys,i)\}} \; ; \; K(sys) \in \mathbb{R} \wedge K(sys) \in [-1,1]$$

where $R(i)$ is the total number of known answers to the question i that are correct and distinct; *answered(sys,i)* is the number of answers given by the system *sys* for the question i; $score(r)$ is the confidence score assigned by the system to the answer r; $eval(r)$ depends on the judgement given by a human assessor.

$$eval(r) = \begin{cases} 1 & \text{if } r \text{ is judged as correct} \\ 0 & \text{if } r \text{ is a repeated answer} \\ -1 & \text{if } r \text{ is judged as incorrect} \end{cases}$$

When $K(sys)$ equals 0 it matches with a system without knowledge that assigns 0 to the confidence score of all their answers. Therefore, $K(sys) = 0$ is established as a baseline and K-measure gives an idea about the system's knowledge.

The answers finding process, accomplished by human assessors, is strongly determined by the evaluation measure. In the case of K-measure the parameter *R(i)* requires a knowledge of all the correct and distinct answers contained in the corpus for each question. This fact introduces a very high cost in the pre-assessment process because it is not easy to ensure that, even with a

human search, all distinct answers for each question have been found in a very large corpus. One alternative is to relax the pre-assessment process and consider only the set of different answers found by humans or systems along the process. Another alternative is to request only one answer per question and ignore recall.

3.2 The *K1*-measure

A second measure, derived from the K-measure, is proposed to evaluate exercises when just one answer per question is requested (number of questions equals number of answers) or when the achievement of all the possible answers by the system is not outstanding for the exercise. That measure has been called *K1*-measure (K-measure for systems giving 1 answer per question) and it is defined as follows:

$$K1(sys) = \frac{\sum_{r \in answers(sys)} score(r) \cdot eval(r)}{\#questions} \; ; \; K1(sys) \in \mathbb{R} \wedge K1(sys) \in [-1, 1]$$

where $score\,(r)$ is the confidence score assigned by the system to the answer r and $eval\,(r)$ depends on the judgement given by a human assessor.

$$eval\,(r) = \begin{cases} 1 & \text{if } r \text{ is judged as correct} \\ -1 & \text{in other case} \end{cases}$$

Again, $K1\,(sys) = 0$ is established as a baseline.

3.3 Comparison with Precedent Measures

Comparing K and *K1* measures with other measures used in precedent QA evaluation exercises, the following differences and similarities have been found:

- **Accuracy measure**, commonly used in all evaluations [1][2][3][7][8][9][10][11], measures the precision in giving correct answers. But it does not take into account the confidence score, as in K and *K1* measures, nor the recall when more than one answer per question is given, as in F-measure or K-measure.

- **Mean F-measure**, used in the QA Track at TREC 2003 [11] and in the QA Challenge at NTCIR 2002 [1], gives a combination between precision and recall, generally the mean of both. As the K-measure, it is designed for systems that must give all the correct answers existing in the corpus for every question. The K-measure takes into account a combination of precision and recall by means of the $max\{R(i), answered(sys, i)\}$ denominator. In addition, K and *K1* measures include the confidence score into their calculations.

- **Mean Reciprocal Rank**, used in the QA Track at TREC [7][8][9][10], in the QA Challenge at NTCIR 2002 [1] and in the QA Track at CLEF 2003 [2] [3]. It is designed for systems that give one or more answers per question, in a decreasing order of confidence. It rewards systems assigning a higher confidence to the correct answers. However, Mean Reciprocal Rank cannot evaluate systems that find several different and correct answers for the same question, and the incorrect answers are not considered as a worse case than the absence of answers.

- **Confident-Weighted Score** (CWS), used in the QA Track at TREC 2002 [10] and in the QA Track at CLEF 2004 [6] as a secondary measure. It is designed for systems that give only one answer per question. Answers are in a decreasing order of confidence and CWS rewards systems that give correct answers at the top of the ranking. Hence, correct answers in the lower zone of the ranking make a very poor contribution to the global valuation, and this contribution is determined by the ranking position instead of the system's self-scoring.

3.4 Correlation Between Self-Scoring and Correctness

Since the confidence score has been included in the K-measure, a high correlation between self-scoring and correctness is expected to produce higher values of K. However, it is interesting to know separately the quality of the scoring given by every system. Hence, it is proposed the use of the correlation coefficient (r) between self-scoring value (in range [0,1]) and the value associated to the human assessment: 1 for the correct answers and 0 otherwise. That is:

$$r(sys) = \frac{\sigma_{assess(sys)score(sys)}}{\sigma_{assess(sys)} \cdot \sigma_{score(sys)}} \; ; \; r(sys) \in \mathbb{R} \wedge r(sys) \in [-1, 1]$$

where $assess(sys)$ and $score(sys)$ are the two multidimensional variables containing the values of the human assessment and the confidence score for the system sys; $\sigma_{assess(sys)}$, $\sigma_{score(sys)}$ are the typical deviations for $assess(sys)$ and $score(sys)$; $\sigma_{assess(sys)score(sys)}$ is the covariance between the two variables.

When a system assigns a $score = 1$ to its correct answers and $score = 0$ to the rest, it obtains a correlation coefficient $r = 1$, meaning that such a system has a perfect knowledge about the correctness of its response. A correlation coefficient equal to 0 indicates that score and correctness have no correlation. A negative value indicates that there is a certain correlation but in the other direction.

3.5 A Case of Study

In the QA 2004 Main Track [6], the confidence score has been requested in order to calculate the CWS as a secondary evaluation measure. This confidence score, together with the human assessments of all the submitted runs, permitted to study the effect of the *K1*-measure in the ranking of systems, and to compare the official measures with this one. No conclusions should be stated about the quality of systems because they should not be compared across different target languages, and also because they did not develop any strategy in order to obtain good values of *K1*.

Table 1 shows the number of given correct answers, CWS, *K1* and the correlation coefficient for all the systems participating in the QA at CLEF 2004 Main Track.

A higher correlation coefficient (higher score for the correct answers) brings associated better values of *K1* for the same or similar number of given correct answers. For example, *ptue041ptpt* ($r > 0.5$) has the 12th position in the ranking of given correct answers and reaches the 1st position for *K1*.

On the contrary, there are some interesting examples, as *fuha041dede* or *dfki041deen*, that have a low or even negative correlation coefficient and experiment a huge drop in the ranking of *K1*.

However, these systems obtain a very good CWS value, showing that CWS does not reward a good correlation between self-scoring and correctness. Why do these systems obtain good values of CWS? The reason can be found when looking at their answers in detail: they tune their score to obtain a better CWS and, obviously, not a better *K1*. For example, when they have not enough confidence in the answer, they return NIL with a score 1, ensuring 20 correct answers (the 20 NIL questions) very high weighted in the CWS measure. All wrong NIL answers (149, with score 1) affect negatively the correlation coefficient and also the *K1*-measure. Adopting a *K1* oriented strategy, they would obtain very good results. For example, if all NIL answers of *fuha041dede* had a score equal to 0 then the correlation coefficient would have been very high ($r = 0.7385$) and the system would have obtained again the first place in the ranking with $K1 = 0.218$.

These systems are an example of how, with the current state-of-the-art, systems can give a very accurate self-scoring.

Since *K1* depends on the number of correct given answers, a good correlation coefficient is not enough to obtain good results: the more correct answers given, the more quantity of positive components conforming the global calculation of *K1*. For example, to beat *fuha041dede* using the mentioned *K1*-oriented strategy ($K1 = 0.218$), a system with perfect scoring (r=1) would need to answer correctly more than 40 questions.

Table 1: **Values and rankings for accuracy, CWS, K1, and correlation coefficient r, for all runs submitted to the Main QA Track at CLEF 2004**

	given correct answers			CWS		*K1*		
run	**#**	**%**	**ranking**	**value**	**ranking**	**value**	**ranking**	**r**
uams042nlnl	91	45.50	1	0.3262	2	0.0078	2	0.1148
uams041nlnl	88	44	2	0.2841	3	0.0063	3	0.0987
uams041ennl	70	35	3	0.2222	4	0.0055	4	0.1105
fuha041dede	67	33.50	4	0.3284	1	-0.3271	27	0.0094
aliv042eses	65	32.50	5	0.1449	8	-0.0416	15	0.1711
aliv041eses	63	31.50	6	0.1218	9	-0.0500	16	0.1099
irst041itit	56	28	7	0.1556	7	-0.1853	19	0.2128
talp042eses	52	26	8	0.1029	12	-0.2252	20	-0.0366
dfki041dede	51	25.50	9..10	N/A †	N/A	0	5..14	N/A
ilcp041itit	51	25.50	9..10	N/A	N/A	0	5..14	N/A
talp041eses	48	24	11	0.0878	15	-0.2464	22	-0.0483
ptue041ptpt	47	23.62	12	0.2162	5	0.0201	1	0.5169
dfki041deen	47	23.50	13	0.1771	6	-0.5131	45	-0.0453
inao041eses	45	22.50	14..15	N/A	N/A	0	5..14	N/A
irst041iten	45	22.50	14..15	0.1215	10	-0.2310	21	0.1411
irst042itit	44	22	16	0.1075	11	-0.3248	26	-0.0188
gine042frfr	42	21	17	0.0954	13	-0.3152	24	0.1917
edin042fren	40	20	18	0.0589	21	-0.4066	38	0.0004
lire042fren	39	19.50	19	0.0754	16	-0.1738	18	0.3707
dltg041fren	38	19	20	N/A	N/A	0	5..14	N/A
inao042eses	37	18.50	21	N/A	N/A	0	5..14	N/A
irst042iten	35	17.50	22	0.0751	17	-0.3300	28	0.0566
edin042deen	34	17	23	0.0527	25	-0.3556	30	0.1124
edin041fren	33	16.50	24	0.0570	22	-0.5336	46	-0.0560
gine042defr	32	16	25	0.0878	14	-0.3009	23	0.3040
gine042esfr	30	15	26	0.0635	19	-0.3757	32	0.1568
dltg042fren	29	14.50	27	N/A	N/A	0	5..14	N/A
edin041deen	28	14	28	0.0492	29	-0.5515	47	-0.0077
gine041defr	27	13.50	29..30	0.0714	18	-0.3945	34	0.2039
gine042itfr	27	13.50	29..30	0.0525	26	-0.4035	37	0.1361
bgas041bgen	26	13	31..33	0.0564	23	-0.3618	31	0.2023
gine041frfr	26	13	31..33	0.0470	32	-0.4523	40	0.1447
gine042nlfr	26	13	31..33	0.0607	20	-0.3884	33	0.1958
gine041esfr	25	12.50	34..36	0.0541	24	-0.4585	41	0.1051
gine042enfr	25	12.50	34..36	0.0481	30	-0.3306	29	0.2462
gine042ptfr	25	12.50	34..36	0.0508	28	-0.4028	36	0.1646
gine041itfr	23	11.50	37	0.0475	31	-0.4013	35	0.1262
sfnx042ptpt	22	11.06	38	N/A	N/A	0	5..14	N/A
cole041eses	22	11	39..41	N/A	N/A	0	5..14	N/A
gine041ptfr	22	11	39..41	0.0413	35	-0.4596	42	0.0970
lire041fren	22	11	39..41	0.0330	37	-0.3200	25	0.2625
hels041fien	21	10.61	42	0.0443	33	-0.1136	17	0.0359
mira041eses	18	9	43	N/A	N/A	0	5..14	N/A
gine041nlfr	17	8.50	44	0.0416	34	-0.4640	43	0.1850
gine041enfr	16	8	45	0.0313	38	-0.4511	39	0.1444
sfnx041ptpt	14	7.04	46	N/A	N/A	0	5..14	N/A
gine041bgfr	13	6.50	47..48	0.0514	27	-0.5603	48	0.1067
gine042bgfr	13	6.50	47..48	0.0380	36	-0.4945	44	0.0928

†**CWS** and **r** are Not Available because 0 was given as confident score for every answer.

4 Results of the Pilot Task

The data from the assessment process for the Pilot Task are shown in Table 2. Only one run from the University of Alicante (UA) [5] was submitted and, therefore, a comparison with other participants cannot be done. The UA system is based in the splitting of nested questions in order to answer questions with temporal restrictions. They have evaluated their system over the TERQAS corpus [4], obtaining better results than in this Pilot Task at CLEF 2004.

Table 2: **Results of the assessment process for the Pilot Task at CLEF 2004.** Data from the run of the University of Alicante.

		# quest.	# known distinct answers	# given answers	questions with at least 1 correct answer	# given correct answers	recall	precision	*K*	r
Definition		2	3	2	0 (0%)	0	0%	0	0	N/A †
Factoid		18	26	42	4 (22.22%)	5	19.23%	11.9%	-0.029	-0.089
List		20	191	55	4 (20%)	6	3.14%	10.9%	-0.070	0.284
Temp.	**Date**	20	20	30	2 (10%)	2	10%	6.67%	-0.019	N/A
	Event	20	20	42	2 (10%)	2	10%	4.76%	-0.024	0.255
	Period	20	20	29	3 (15%)	3	15%	10.3%	-0.003	0.648
Total		100	280	200	15 (15%)	18	6.43%	9%	-0.086	0.246

†**r** is Not Available because 0 was given for every component of any variable.

The UA system has correctly answered 15% of the questions. The best result corresponds to factoid questions with a 22.22% of questions with a correct answer. However, in the past edition of QA at CLEF, this team obtained better results (up to 40% of questions with a correct answer) [2]. This results show that the questions posed in the Pilot Task have been too difficult.

The UA system never gave more than three answers per question, independently of the type of formulated question. It seems an heuristically established limit for the system that has affected the achievement of good conjunctive and disjunctive list answers.

41 questions got NIL as an answer, with a confidence score of 0 for all them. Unfortunately, these 41 questions had at least one answer in the corpus. On the other hand, the UA system did not identify the 2 posed NIL questions.

Finally, it seems that the UA system did not play with the score value in the best way. The maximum value given for the confidence score was 0.5002 and several questions with only one correct answer in the corpus had associated several different answers with similar confidence score. The K-measure for the UA's exercise was $K = -0.086$ with a correlation coefficient of $r = 0.246$ between self-scoring and real assessment.

5 Conclusions and Future Work

Questions whose answer is a conjunctive or a disjunctive list, and questions with temporal restrictions, still remain a challenge for most QA systems. However, these are only a few types of *difficult* questions which QA systems will have to manage in the near future. A specialization and further collaboration among teams could be expected in order to achieve QA systems with higher accuracy and coverage for different types of questions. In fact, the QA Main Track at CLEF shows that different participant systems answer correctly different subsets of questions.

Two measures have been proposed in order to reward systems that give a confidence score with a high correlation with human assessments and, at the same time, return more correct answers and less incorrect ones. The case of study shows that systems are able to give very accurate self-scoring, and that the K and *K1* measures reward it. However, systems don't need to respond all the questions to obtain good results, but to find a good balance between the number of correct answers and the accuracy of their confidence score.

On the one hand, this seems a good way to promote the development of more accurate systems with better answer validation. On the other hand, it is a good way to permit some specialization, to open the possibility of posing new types of questions and, at the same time, to leave the door open for new teams starting to develop their own systems.

6 Acknowledgements

This work has been partially supported by the Spanish Ministry of Science and Technology within the following projects: TIC-2002-10597-E Organization of a Competitive Task for QA Systems; TIC-2003-07158-104-01 Answer Retrieval from Digital Documents, R2D2; and TIC-2003-07158-C04-02 Multilingual Answer Retrieval Systems and Evaluation, SyEMBRA.

We are grateful to Julio Gonzalo, from UNED-NLP Group, and Alessandro Vallin, from ITC-Irst (Italy), for their contributions to this work. In addition, we would like to thank the University of Alicante team for their effort in participating in the Pilot Task.

References

[1] J. Fukumoto, T. Kato, and F. Masui. Question Answering Challenge (QAC-1). An Evaluation of Question Answering Task at NTCIR Workshop 3. In Keizo Oyama, Emi Ishida, and Noriko Kando, editors, *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*. National Institute of Informatics, 2003.

[2] B. Magnini, S. Romagnoli, A. Vallin, J. Herrera, A. Peñas, V. Peinado, F. Verdejo, and M. de Rijke. The Multiple Language Question Answering Track at CLEF 2003. In C. Peters, J. Gonzalo, M. Braschler, and M. Kluck, editors, *Comparative Evaluation of Multilingual Information Access Systems. Results of the CLEF 2003 Evaluation Campaign*, volume 3237 of *LNCS*, pages 479–495. Springer-Verlag, 2004.

[3] A. Peñas, J. Herrera, and F. Verdejo. Spanish Question Answering Evaluation. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing, CICLing 2004*, volume 2945 of *LNCS*, pages 472–483. Springer-Verlag, 2004.

[4] J. Pustejovsky et al. TERQAS Final Report. Technical report, MITRE, http://www.cs.brandeis.edu/~jamesp/arda/time/readings.html, October 2002.

[5] E. Saquete, P. Martínez-Barco, R. Muñoz, and J.L. Vicedo. Splitting complex temporal questions for question answering systems. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 566–573, Barcelona, Spain, July 2004.

[6] A. Vallin et al. Overview of the CLEF 2004 Multilingual Question Answering Track. In *Proceedings of the CLEF 2004 Workshop*, Bath, United Kingdom, September 2004.

[7] E. M. Voorhees. The TREC-8 Question Answering Track Report. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of the Eigthh Text REtrieval Conference (TREC 8)*, volume 500-246 of *NIST Special Publication*, pages 77–82, 1999.

[8] E. M. Voorhees. Overview of the TREC-9 Question Answering Track. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of the Ninth Text REtrieval Conference (TREC 9)*, volume 500-249 of *NIST Special Publication*, pages 71–79, 2000.

[9] E. M. Voorhees. Overview of the TREC 2001 Question Answering Track. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*, volume 500-250 of *NIST Special Publication*, pages 42–51, 2001.

[10] E. M. Voorhees. Overview of the TREC 2002 Question Answering Track. In E. M. Voorhees and L. P. Buckland, editors, *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, volume 500-251 of *NIST Special Publication*, 2002.

[11] E. M. Voorhees. Overview of the TREC 2003 Question Answering Track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, volume 500-255 of *NIST Special Publication*, pages 54–68, 2003.