Overview of the CLEF 2006 Multilingual Question Answering Track

Bernardo Magnini[1], Danilo Giampiccolo[2], Pamela Forner[2], Christelle Ayache[3], Valentin Jijkoun[4], Petya Osenova[5], Anselmo Peñas[6], Paulo Rocha[7], Bogdan Sacaleanu[8], and Richard Sutcliffe[9]

[1] FBK-irst, Trento, Italy
magnini@itc.it
[2] CELCT, Trento, Italy
{giampiccolo, forner}@celct.it
[3] ELDA/ELRA, Paris, France
ayache@elda.fr
[4] Informatics Institute, University of Amsterdam, The Netherlands
jijkoun@science.uva.nl
[5] BTB, Bulgaria
petya@bultreebank.org
[6] Departamento de Lenguajes y Sistemas Informáticos, UNED, Madrid, Spain
anselmo@lsi.uned.es
[7] Linguateca, SINTEF ICT, Norway and Portugal
Paulo.Rocha@alfa.di.uminho.pt
[8] DFKI, Germany
Bogdan.Sacaleanu@dfki.de
[9] DLTG, University of Limerick, Ireland
richard.sutcliffe@ul.ie

Abstract. Having being proposed for the fourth time, the QA at CLEF track has confirmed a still raising interest from the research community, recording a constant increase both in the number of participants and submissions. In 2006, two pilot tasks, WiQA and AVE, were proposed beside the main tasks, representing two promising experiments for the future of QA.Also in the main task some significant innovations were introduced, namely list questions and requiring text snippet(s) to support the exact answers. Although this had an impact on the work load of the organizers both to prepare the question sets and especially to evaluate the submitted runs, it had no significant influence on the performance of the systems, which registered a higher Best accuracy than in the previous campaign, both in monolingual and bilingual tasks. In this paper the preparation of the test set and the evaluation process are described, together with a detailed presentation of the results for each of the languages. The pilot tasks WiQA and AVE will be presented in dedicated articles.

1 Introduction

Inspired by previous TREC evaluation campaigns, QA tracks have been proposed at CLEF since 2003. During these years, the effort of the organisers has been focused on two main issues. One issue was to offer an evaluation exercise characterised by cross-linguality, covering as many languages as possible. From this perspective, major

C. Peters et al. (Eds.): CLEF 2006, LNCS 4730, pp. 223 – 256, 2007.

attention has been given to European languages, adding at least one new language each year, but keeping the offer open to languages from all-over the world, as the use of Indonesian shows. The other important issue was to maintain a balance between the established procedure inherited from the TREC campaigns and innovation. This allowed newcomers to join the competition and, at the same time, offered "veterans" more challenges. Following these principles, in QA@CLEF 2006 two pilot tasks, namely WiQA and Answer Validation Exercise (AVE), were proposed together with a main task. As far as the latter is concerned, the most significant innovation was the introduction of LIST questions, which had also been considered for previous competitions, but had previously been avoided due to the problems that their selection and assessment implied.

Other important innovations consisted in the possibility to return more than one answer per question, and by the request to provide text snippets together with the docid to support the exact answer. All these changes implied the necessity of introducing new evaluation measures, which would account also for List and multiple answers. Nevertheless, the evaluation process proved to be more complicated than expected, partly because of the excessive workload that multiple answers represented for groups already in charge for a larger number of runs. As a consequence, some groups, like the Spanish and the English ones, could only correct one answer per question, which decreased the possibility of comparisons between runs.

As a general remark, it can be said that the positive trend in participation registered in the previous campaigns was confirmed, and 10 new participants joined the competition from Europe, Asia and America.

As reflected in the results, systems' performance improved considerably, with the Best Accuracy increasing from 64% to 68% in the monolingual tasks, and, more significantly, from 39% to 49% in the bilingual ones.

This paper describes the preparation process and presents the results of the QA track at CLEF 2006. In section 2, the task is described in detail. The different phases of the Gold Standard preparation are exposed in section 3. After a quick presentation of the participants in section 4, the evaluation procedure and the results are reported respectively in section 5 and 6. In section 7, some final considerations are given about this campaign and the future of QA@CLEF.

2 Tasks

In 2006 campaign, the procedure consolidated in previous competitions was used. Accordingly, there was a main task (which was comprehensive of a monolingual task and several cross-language sub-tasks), and two pilot tasks described below:

1. **WiQA:** developed by Maarten de Rijke. The purpose of the WiQA pilot is to see how IR and NLP techniques can be effectively used to help readers and authors of Wikipages access information spread throughout Wikipedia rather than stored locally on the pages.[2]
2. **Answer Validation Exercise (AVE):** A voluntary exercise to promote the development and evaluation of subsystems aimed at validating the correctness of the answers given by a QA system. The basic idea is that once

> a pair [answer + snippet] is returned by a QA system, a hypothesis is built by turning the pair [question + answer] into the affirmative form. If the related text (a snippet or a document) semantically entails this hypothesis, then the answer is expected to be correct [3].

Two specific papers in the present Working Notes are dedicated to these pilot tasks. More detailed information, together with the results, can be found there.

In addition to the tasks proposed during the actual competition, a "time-constrained" QA exercise was proposed by the University of Alicante during the CLEF 2006 Workshop. In order to evaluate the ability of QA systems to retrieve answers in real time, the participants were given a time limit (e.g. one or two hours) in which to answer a set of questions. These question sets were different and smaller than those provided in the main task (e.g. 15-25 questions). The initiative was aimed towards providing a more realistic scenario for a QA exercise.

The main task was basically the same as in previous campaigns. Some new ideas were implemented in order to make the competition more challenging. The participating systems were fed a set of 200 questions, which could be about:

- facts or events (F-actoid questions);
- definitions of people, things or organisations (D-efinition questions);
- lists of people, objects or data (L-ist questions).

The systems were then asked to return from one to ten exact answers. “Exact” meant that neither more nor less than the information required is given. The answer needed to be supported by the docid of the document(s) in which the exact answer was found, and by one to ten text snippets which gave the actual context of it.

The text snippets were to be put one next to the other, separated by a tab. The snippets were substrings of the specified documents. They should provide enough context to justify the exact answer suggested. Snippets for a given response had to be a set of sentences of not more than 500 bytes in total (although for example the Portuguese group accepted – and actually preferred – length to be specified in sentences). There were no particular restrictions on the length of an answer-string, but unnecessary pieces of information were penalized, since the answer was marked as ineXact. Since Definition questions may have long strings as answers, they were (subjectively) assessed mainly on their informativity and usefulness, and not on exactness. The tasks were both:

- monolingual, where the language of the question (Source language) and the language of the news collection (Target language) were the same;
- cross-lingual, where the questions were formulated in a language different from that of the news collection.

Eleven source languages were considered, namely, Bulgarian, Dutch , English, French, German, Indonesian, Italian, Polish , Portuguese, Romanian and Spanish. Note the loss of Finnish, and the introduction of Polish and Romanian with respect to last year. All these languages were also considered as target languages, except for Indonesian, Polish and Romanian. These three languages had no news collection available for the queries. As was done for Indonesian in the previous two campaigns, the English question set was translated into Indonesian (IN), Polish (PL) and

Table 1. Task activated in 2006

SOURCE LANGUAGES (questions)	TARGET LANGUAGES (corpus and answers)							
	BG	DE	EN	ES	FR	IT	NL	PT
BG								
DE								
EN								
ES								
FR								
IN								
IT								
NL								
PL								
PT								
RO								

Romanian (RO), and the German question set into Romanian (RO). Only the bilingual tasks IN-EN, PL-EN, RO-EN and RO-DE were activated. In the case of IN-EN, PL-EN, and RO-EN, the questions were posed in the respective language (i.e. IN, PL, RO), while the answers were retrieved from the English collection. In the RO-DE case, the question was made in Romanian, whilst the answer was retrieved from the German collection.

As shown in Table 1, 24 tasks were proposed and divided in:

- 7 Monolingual -i.e. Bulgarian (BG), German (DE), Spanish (ES), French (FR), Italian (IT), Dutch (NL), and Portuguese (PT);
- 17 Cross-lingual.

As customary in recent campaigns, a monolingual English (EN) task was not available as it seems to have been already thoroughly investigated in TREC campaigns, even though English was both source and target language in the cross-language tasks.

Although the task was not radically changed with regard to previous campaigns, some new elements were introduced. The most important one was the addition of List questions to the question sets, which implied some major issues. For this first year of QA@CLEF, we were not too strict on the definition of lists, using both questions asking for a specific finite number of answers (that could be called "closed lists") e.g.:

Q: What are the names of the two lovers from Verona separated by family issues in one of Shakespeare's plays?
A: Romeo and Juliet.

and open lists ,where as many correct answers could be returned, e.g.

Q: Name books by Jules Verne.

Table 2. Tasks chosen by at least 1 participant in QA@CLEF campaigns

	MONOLINGUAL	CROSS-LINGUAL
CLEF 2003		
CLEF 2004	**6**	**13**
CLEF 2005	**8**	**15**
CLEF 2006	**7**	**17**

and let organizing groups decide on how to assess the answers to these different kinds of questions.

Other innovations were:

- the input format, where the type of question (F,D,L) was no longer indicated;
- and the result format, where up to a maximum of ten answers per question was allowed, with one to ten text snippets supporting the exact answer.

3 Test Set Preparation

Following the procedure established in previous campaigns, initially each organising group (one for each Target language) was assigned a number of topics taken from the CLEF IR track on which candidates' questions were based. This choice was originally made to reduce the number of duplicates in the multilingual question set. As the number of new topics introduced in 2006 was small, old topics were simply reassigned to different groups. Some groups questioned this methodology, preferring to produce questions with other methods instead of following particular topics. The topics, and hence the questions, were aimed at data collections composed of news articles provided by ELRA/ELDA dating back to 1994/1995; with the exception of Bulgarian, which dated back to 2000 (see Table 3). The choice of a different collection was a matter for long discussion, copyright issues remaining a major obstacle. A step towards a possible solution was nevertheless made by the proposal of the WiQA pilot task, which represents a first attempt to set the QA competitions in their natural context, i.e. the Internet.

Initially, 100 questions were selected in each of the source languages, distributed between Factoid, Definition and List questions.

Factoid questions are fact-based questions, asking for the name of a person, a location, the extent of something, the day on which something happened, etc.

Table 3. Document collections used in CLEF 2006

TARGET LANG..	COLLECTION	PERIOD	SIZE
Bulgarian (BG)	Sega	2002	120 MB (33,356 docs)
	Standart	2002	93 MB (35,839 docs)
Germany (DE)	Frankfurter Rundschau	1994	320 MB (139,715 docs)
	Der Spiegel	1994/1995	63 MB (13,979 docs)
	German SDA	1994	144 MB (71,677 docs)
	German SDA	1995	141 MB (69,438 docs)
English (EN)	Los Angeles Times	1994	425 MB (113,005 docs)
	Glasgow Herald	1995	154 MB (56,472 docs)
Spanish (ES)	EFE	1994	509 MB (215,738 docs)
	EFE	1995	577 MB (238,307 docs)
French (FR)	Le Monde	1994	157 MB (44,013 docs)
	Le Monde	1995	156 MB (47,646 docs)
	French SDA	1994	86 MB (43,178 docs)
	French SDA	1995	88 MB (42,615 docs)
Italian (IT)	La Stampa	1994	193 MB (58,051 docs)
	Itallian SDA	1994	85 MB (50,527 docs)
	Itallian SDA	1995	85 MB (50,527 docs)
Dutch (NL)	NRC Handelsblad	1994/1995	299 MB (84,121 docs)
	Algemeen Dagblad	1994/1995	241 MB (106,483 docs)
Portuguese (PT)	Público	1994	164 MB (51,751 docs)
	Público	1995	176 MB (55,070 docs)
	Folha de São Paulo	1994	108 MB (51,875 docs)
	Folha de São Paulo	1995	116 MB (52,038 docs)

The following 6 answer types for factoids were considered:

- PERSON (e.g. "*Who was Lisa Marie Presley's father*?")
- TIME (e.g. "*What year did the Second World War finish?*")
- LOCATION (e.g. "*What is the capital of Japan?*")

- ORGANIZATION (e.g. "*What party did Hitler belong to?*")
- MEASURE (e.g. "*How many monotheistic religions are there in the world?*")
- OTHER, i.e. everything else that does not fit into the other five categories (e.g. "*What is the most-read Italian daily newspaper?*")

Definition questions, i.e. questions like "What/Who is X?", were divided into the following categories:

- PERSON -i.e. questions asking for the role, job, and/or important information about someone (e.g. "*Who is Lisa Marie Presley?*");
- ORGANIZATION -i.e. questions asking for the mission, full name, and/or important information about an organization (e.g. "*What is Amnesty International?*" or "*What is the FDA*?");
- OBJECT -i.e. questions asking for the description or function of objects (e.g. "*What is a Swiss army knife?*", "*What is a router?*");
- OTHER -i.e. question asking for the description of natural phenomena, technologies, legal procedures etc. (e.g. "*What is a tsunami?*", "*What is DSL?*", "*What is impeachment?*").

The last two categories were especially added to reduce the numbers of definition questions which may be answered very easily (such as acronyms concerning organizations, which are usually answered rendering the abbreviation in full, and people's job-description, which are usually found as appositions of proper names in news text).

As mentioned above, questions that require a list of items as answers, were introduced for the first time. (e.g. *Name works by Tolstoy*).

Among these three categories, a number of NIL questions, i.e. questions that do not have any known answer in the target document collection, were distributed. They are important because a good QA system should identify them, instead of returning wrong answers.

Table 4. Test set breakdown according to question type

	F (150)	**D (40)**	**L (10)**	**T (40)**	**NIL (20)**
BG	145	43	12	26	17
DE	153	37	10	45	20
EN	150	40	10	40	18
ES	148	42	10	40	21
FR	148	42	10	40	20
IT	147	41	12	38	20
NL	147	40	13	30	20
PT	143	47	9	23	18

Three different types of temporal restriction – a temporal specification that provides important information for the retrieval of the correct answer, were associated to a certain number of F, D, L, more specifically:

- restriction by DATE (e.g. "*Who was the US president in 1962?"; "Who was Berlusconi in 1994?")*
- restriction by PERIOD (e.g. "*How many cars were sold in Spain between 1980 and 1995?")*
- restriction by EVENT (e.g. "*Where did Michael Milken study before enrolling in the university of Pennsylvania?*")

The distribution of the questions among these categories is described in Table 4.

Each of the question sets was then translated into English, so that each group could choose additional 100 questions from those proposed by the others and translate them in their own languages. At the end, each source language had 200 questions, which were collected in an XML document. Unlike in the previous campaigns, the questions were not translated in all the languages due to time constraints, and the Gold Standard contained questions in multiple languages only for activated tasks. Since Indonesian, and Romanian did not have a data collection of their own, the English question set was translated, so that the cross-lingual subtasks IN-EN and RO-EN were made available. As not all questions had been previously translated, a translation of the target language question sets into the source languages was needed for cross-language sub-tasks which had at least one registered participant.

4 Participants

The number of participants has constantly grown over the years [see Table 5]. In fact, about ten new groups have joined the competition each year, and in 2006 a total of 30 participants was reached.

Table 5. Number of participating groups

	America	Europe	Asia	Australia	TOTAL	Registered participants	New groups	Veterans	Absent veterans
CLEF 2003	**3**		**-**	**-**	**8**				
CLEF 2004	**1**	**17**	**-**	**-**	**18 (+125%)**				
CLEF 2005	**1**	**22**	**1**	**-**	**24(+33%)**	**27**	**9**	**15**	**4**
CLEF 2006	**4**	**24**	**2**	**-**	**30 (+25%)**	**36**	**10**	**20**	**4**

For the record, the number of groups which registered for the competition but did not actually participate in it was six, while four groups which took part in QA2005 did not show up in 2006. From a geographical perspective, most groups came from

Europe, but in 2006 there was an increase in participants from both Asia and America [see Table 5].

The increase in the number of submitted runs corresponded to that of the participants. Of higher significance is the slight decrease registered in monolingual subtasks to the advantage of bilingual ones. This indicates that QA@CLEF is becoming increasingly cross-lingual, as it was originally set out to be.

Table 6. Number of submitted runs

	Number of submitted runs	Monolingual	Cross-lingual
CLEF 2003	**17**	**6**	**11**
CLEF 2004	**48**	**20**	**28**
CLEF 2005	**67**	**43**	**24**
CLEF 2006	**77**	**42**	**35**

5 Evaluation

The introduction of list questions, the possibility to return multiple answers, and the requirement of supporting the answers with snippets of texts from the relevant documents made the evaluation process more difficult. Moreover, in some languages the large amount of data requiring assessment made it impossible for the judging panels to correct more than one answer per question. Therefore, only the first answers were evaluated in runs that had English and Spanish as a target. In all other cases at least the first three answers were evaluated.

Considering these issues, it was decided to follow the procedure utilised during the previous campaign. The files submitted by the participants in all tasks were manually judged by native speakers. Each language coordination group guaranteed the evaluation of at least one answer per question.

If a group decided to assess more than one answer per question, the answers were assessed in the order they occurred in the submission file and the same number was applied to all questions, and all the runs assessed by the group. The exact answer (i.e. the shortest string of words which is supposed to provide the exact amount of information to answer the question) was assessed as:

- R (Right) if correct;
- W (Wrong) if incorrect;
- X (ineXact) if contained less or more information than that required by the query;
- U (Unsupported) if either the docid was missing or wrong, or the supporting snippet did not contain the exact answer.

Most assessor-groups managed to guarantee a second judgement of all the runs, with a good average inter-assessor agreement. As far as the evaluation measures are concerned, the list questions had to be scored separately, and different groups returned a different number of answers for originally meant Factoid and Definition questions. As a consequence, we decided to provide the following measures:

- accuracy, as the main evaluation score, defined as the average of SCORE(q) over all 200 questions q;
- the mean reciprocal rank (MRR) over N assessed answers per question. That is, the mean of the reciprocal of the rank of the first correct label over all questions;
- the K1 measure used in earlier QA@CLEF campaigns [4]
- the Confidence Weighted Score (CWS) 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 [4].

Although some other kinds of measures have been proposed and used in CLEF 2005, such as a more detailed analysis/breakdown of bad answers by the Portuguese group [7], they were not considered this year. Also, issues like providing more accurate description of what X means: too much or too little were only distinguished by the Portuguese assessors, argued for i.a. in Rocha and Santos [6].

6 Results

As far as accuracy is concerned, a general improvement has been noticed, as Figure 1 shows.

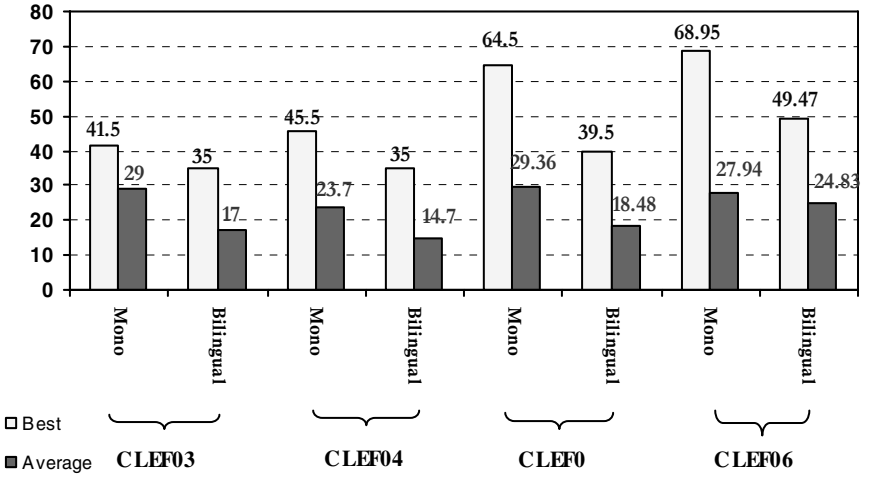

Fig. 1. Best and average scores in CLEF QA campaigns

In detail, Best Accuracy in the monolingual task improved by 6.9%, passing from last year's 64.5% to 68.95%, while Best Accuracy in cross-language tasks passed from 39.5% to 49.47%, recording an increment of 25.2%. As far as average performances are concerned, a slight decrease has been recorded in the monolingual tasks, which went from 29.36% to 27.94%. This probably was due to the number of newcomers which tested their systems for the first time.

As a general remark, best performances has been quite stable, with most languages registering similar or better scores than last campaigns (see Figure 2).

Table 7. Best accuracy scores compared with K1, MRR, and CWS

FILE NAME	OVERALL ACCURACY	K1	MRR	CWS
BEST				
syna061frfr.txt	68.95%	0.2832	0.6895	0.56724
inao061eses.txt	52.63%	0.0716	0.5263	0.43387
ulia061frfr.txt	46.32%	0.0684	0.4632	0.46075
ulia062frfr.txt	45.79%	0.0579	0.4579	0.45546
vein061eses.txt	42.11%	-0.0657	0.4211	0.33582
alia061eses.txt	37.89%	-0.1232	0.3763	0.23630
upv_061eses.txt	36.84%	0.0014	0.3684	0.22530
ulia061enfr.txt	35.26%	-0.1684	0.3526	0.34017

Although also in 2006 campaign self confidence score was not returned by all systems, data about the confidence were plentiful, and allowed to consider the additional evaluation measures, i.e. K1, CWS and MMR. Generally speaking, systems with high accuracy scored accordingly well also with these measures, implying that best systems provide high self confidence, as Table 7 shows.

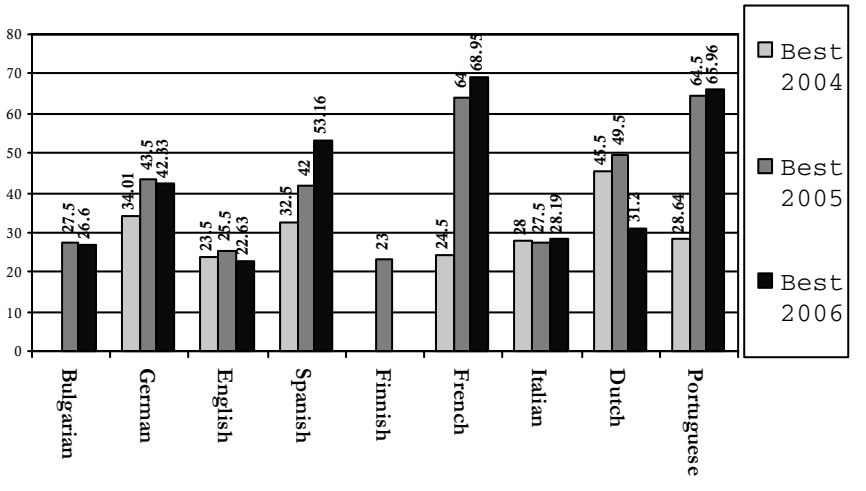

Fig. 2. Best results in 2004, 2005 and 2006

Here below a more detailed analyses of the results in each language follows, giving more specific information on the performances of systems in the single sub-tasks and on the different types of questions, providing the relevant statistics and comments.

6.1 Bulgarian as Target

At CLEF 2006 Bulgarian was addressed as a target language for the second time. This year there was no change in the number of the participants -- again two groups took part in the monolingual evaluation task with Bulgarian as a target language: BTB at Linguistic Modelling Laboratory, Sofia and The Joint Research Centre (JRC), Ispra.

Three runs altogether were submitted – one by the first group and two by the second group with insignificant difference between them. The 2006 results are presented in Table 8 below. First, the correct answers in numbers and percentage are given (Right) per run. Then the wrong (W), inexact (X) and unsupported answers (U) are shown in numbers. Further, the number of the factoids (F), temporally restricted questions (T), definitions (D) and list questions (L) are given. Also, the percentage of the correct answers per each type is registered in Table 8. NIL questions are presented as the number of correctly and wrongly returned answers by the systems with NIL marking. It is obvious that the systems returned NIL answer also when they could not detect a possibly existing answer in the corpus themselves. In our opinion, the present NIL marking might be divided into two labels: NIL = no answer in the corpus is existing and CANNOT = the system itself cannot find an answer. In this way the evaluation would be more realistic. Main reciprocal rank score is provided in the last column of the table.

Table 8a. Results at the Bulgarian as target, monolingual

	Right		W	X	U
Run	#	%	#	#	#
btb061	50	26.60	132	4	1
jrc061	22	11.70	162	4	0
jrc062	22	11.70	160	6	0

As it can be seen, this year the first system performs better. However, its overall accuracy is slightly worse with respect to the 2005 best accuracy result, achieved then by IRST, Trento. Now it is 26.60 %, while in 2005 it was 27.50 %. However, the BTB 2005 year result was significantly improved. Both systems 'crashed' at temporally restricted questions with no single match (see the empty slots in the table). It is a step back from 2005, when both systems had some hits, best of which scored 17.65 %. List questions are also very poorly answered (1 correct answer per run).

The only outperforming results in comparison with the last year are the following: the improvement of the definition type answers (from 42 % to 55.81 %) and the raise of the main reciprocal rank score (from 0.160 to 0.2660).

Table 8b. Results at the Bulgarian as target, monolingual

Run	% F [119]	% T [26]	% D [43]	% L [12]	NIL [16] right	NIL [16] wrong	r
btb061	17.93	-	55.81	0.0833	11	120	0.2660
jrc061	6.90	-	27.91	0.0833	12	155	0.1170
jrc062	6.90	-	27.91	0.0833	13	154	0.1170

The introduction of the snippet support proved out to be a good idea. There was only 1 unsupported answer in all three runs.

The interannotator agreement was very high due to two reasons: first, the number of the answered questions was not very high, and second, there were strict guidelines for the interpretation of the answers, based on our last year experience.

In spite of the somewhat controversial results from the participating systems this year, there is a lot of potential in the task of Bulgarian as a target language in several aspects: investing in the development of the present systems and creating new systems. We hope that Bulgarian will become even more attractive as an EU language.

6.2 Dutch as Target

This year three teams that took part in the CLEF QA track used Dutch as the target language: the University of Amsterdam, the University of Groningen and the University of Roma – 3, with six runs submitted in total: three Dutch monolingual and three crosslingual (English to Dutch). All runs were assessed by two assessors, with the overall inter-assessor agreement 0.96.

Table 9a. Results at the Dutch as target, monolingual

Run	Right #	Right %	W #	X #	U #
Gron061nlnl	58	**31.02**	115	11	3
Isla061nlnl	40	21.39	141	4	2
Isla062nlnl	41	21.93	139	4	3

For creating the gold standard for Dutch, the assessments were automatically reconciled in favour of more lenient assessments: for example, in case the same answer was assessed as W (incorrect) by one assessor and as X (inexact) by another, the X judgement was included in the gold standard. The results of the evaluation of the six runs are provided in Tables 9 and 10. The columns labelled Right, W, X and U give the results for factoid, definition and temporally restricted questions.

Table 9b. Results at the Dutch as target, monolingual

Run	% F [146]	% T [0]	% D [40]	P@N (lists) [13]	Accuracy NIL [10]	MRR [187]
Gron061nlnl	**27.40**	0.00	45.00	**23.08**	0	0.3460
Isla061nlnl	21.23	**0.00**	**22.50**	0.00	0.1346	**0.2341**
Isla062nlnl	21.92	0.00	**22.50**	0.00	0.1346	0.2357

An interesting thing to notice about this year's task is that the overall scores of the systems are lower, compared to the last year's numbers (44% and 50% of correct answers to factoid questions last year).

Table 10a. Results at the Dutch as target, cross-lingual (English to Dutch)

Run	Right		W	X	U
	#	%	#	#	#
Gron061ennl	38	**20.32**	139	7	3
Roma061ennl	25	13.37	150	6	3
Roma062ennl	25	13.37	149	7	3

This year's questions were created by annotators who were explicitly instructed to think of "harder" questions, that is, involving paraphrases and some limited general knowledge reasoning. It would be interesting to compare the performance of this year's systems on last year's questions to the previous results of the campaign.

Table 10b. Results at the Dutch as target, cross-lingual (English to Dutch)

% F [146]	% T [0]	% D [40]	P@N (lists) [13]	Accuracy NIL [10]	MRR [187]
18.37	0.00	28.21	**6.15**	0.1481	0.2239
11.56	**0.00**	**20.51**	17.95	0.0769	**0.1430**
11.56	0.00	**20.51**	15.38	0.0769	0.1529

6.3 English as Target

Creation of Questions. The question for creation of the questions was very similar to last year and is now a well understood procedure. This year it was required to store supporting snippets for the reference answers but this was not difficult and is well worth the trouble. As previously, we were requested to set Temporarily Restricted questions and to distribute these in a prescribed way over the various Factoid question types (PERSON, LOCATION etc). We achieved our quotas but this was extremely difficult to accomplish and we do not feel the time spent is worthwhile as the addition of temporal restrictions more than doubles the time taken to generate the questions.

On the other hand, as the restrictions are frequently synthetic in nature, our knowledge of how to solve these important questions does not necessarily advance from year to year.

Searching for Definition questions (or indeed any questions beyond Factoids) is always very interesting work but the method of evaluation was not clarified this year. So, while the topics we selected do follow the guidelines, we were not required to (or indeed able to) state at generation time exactly what a complete and correct answer should look like. In consequence we can not conclude much from an analysis of the answers returned by systems to such questions.

Summary Statistics for all the Runs. Overall, thirteen cross-lingual runs with English as a target were submitted. The results are shown in. Ten groups participated in seven languages, French, German, Indonesian, Italian, Romanian, Polish and Spanish.[1] There were three groups for French, two for Spanish and one for all the rest

Results Analysis. There were three main types of question this year, Factoids, Definitions and Lists and we consider the results over these types as well as considering the best scores overall. The most indicative measure overall is a simple count of correct answers and this is what we have used. For the 150 Factoids the top four systems were lire062fren (39), lire061fren (33), dltg061fren (32) and aliv061esen (29). The other results are not greatly different from last year. The top result of 39/150 amounts to 26%. For the 40 definitions, the picture is similar. The top five results are aliv062esen (11), lire061fren (10), aliv061esen (9), lire062fren (9) and dfki061deen (8). For each of the ten list questions, a system could return up to ten candidate answers. Considering both a simple count of correct answers and the P@N score achieved, the top five results by count are uaic061roen (10, 0.11), lire061fren (9, 0.09), irst061iten (8, 0.16), lire062fren (8, 0.08) and dfki061deen (6, 0.2). The ordering for the P@N score differs: dfkienen (6, 0.2) irst061iten (8, 0.16), uaic061roen (10, 0.11), lire061fren (9, 0.09) and lire062fren (8, 0.08).

Assessment Procedure. This approach to assessment was broadly similar to that of last year. However, as the format of the runs had changed, we decided not to use the NIST software but to work with the bare text files instead. It had been intended to double-judge all the questions but unexpectedly and at the last moment this proved not to be possible due to the absence of an assessor. There were 200 questions in all. One assessor judged all answers to questions 1-100 while the other two judged all answers to questions 101-200.

There were considerable practical problems with the assessment of runs this year. Firstly, several runs used invalid run tags. Secondly two of the runs were answering the questions in a completely different order! Thirdly, one question in these two runs was different from the question being answered by the other systems in that position. Fourthly, one run had the fields in the wrong order. Fifthly one run used NULL instead of NIL while another run used nil. Luckily we spotted problems 2 and 3 and

[1] A Polish-English run -'utjp061plen'- was submitted, achieving an overall accuracy of 86%. However, as the system report did not provide enough information to support such a high score, we have decided not to validate this result.

Table 11. Results of English runs

Run	R #	W #	X #	U #	% F [150]	% D [40]	P@N for L [10]	OVERALL ACCURACY %
aliv061esen	38	142	4	6	19.33	22.50	0.0411	20.00
aliv062esen	29	156	3	2	12.00	27.50	0.0200	15.26
aske061esen	10	134	11	34	6.67	0.00	0	5.26
aske061fren	7	135	10	37	3.33	5.00	0.0100	3.68
dfki061deen	34	147	9	0	17.33	20.00	0.2000	17.89
dltg061fren	36	138	14	2	21.33	10.00	0.2000	18.95
irst061iten	24	152	3	11	16.00	0.00	0.1600	12.63
lire061fren	43	138	2	7	22.00	25.00	0.0900	22.63
lire062fren	48	130	2	10	26.00	22.50	0.0800	25.26
uaic061roen	25	150	7	8	15.33	5.00	0.1131	13.16
uaic062roen	18	171	1	0	12.00	0.00	0.0800	9.47
uind061inen	14	159	4	13	9.33	0.00	0	7.37

were able to correct them and indeed all the others but this was extremely time consuming and difficult.

As in all previous years the runs were anonymised by a third party so none of the assessors knew either the origin of a run or the original source language.

This year it had been decided to allow multiple answers to Factoid and Definition questions (up to ten per question). The rationale for this was never quite clear since the whole objective of Question Answering (as against Information Retrieval) is to return only the right answer. Even in cases where there are genuinely several right answers (a rare situation in our carefully designed question sets) a system should still return a correct answer in the first place. For this reason and due to our limited time and resources, we only judged the first answer returned to Factoid and Definition questions. For List questions, all candidate answers were judged, as is normal at TREC.

For the questions double judged, we measured the agreement level. There were 149 differences over thirteen runs of 100 questions. This amounts to 149/1300 i.e. 11% disagreement or 89% agreement. The overall figure for last year was 93%.

Concerning the judgement process itself, Factoids and Lists did not present a problem as we were very familiar with them. On the other hand Definitions were in the same state as last year in that they had been included in the task without a suitable evaluation procedure having been defined. In consequence we used the same approach as last year: If an answer contained information relevant to the question and also contained no irrelevant information, it was judged R if supported, and U

otherwise. If both relevant and irrelevant information was present it was judged X. Finally, if no relevant information was present, the answer was judged W.

Comment and Conclusions. The number of runs judged (12) was thes same as last year. However, two source languages were introduced: Indonesian and Romanian. The results themselves were also broadly similar .

Definition questions remained in the same unspecified state as previously. This means that we have not been successful in stretching the boundaries of question answering beyond Factoids which are now very well understood. This is a great pity as the extraction of useful 'definition type' information on a topic is a very useful task for groups to study but it is one which needs to be carefully quantified.

The introduction of snippets was very helpful at question generation time and also invaluable for judging the answers. Snippets are a great step forward for CLEF and are the most significant development for the QA Track this year.

6.4 French as Target

This year (as last year) seven groups took part in evaluation tasks using French as target language: four French groups: Laboratoire d'Informatique d'Avignon (LIA), CEA-List, Université de Nantes (LINA) and Synapse Développement; one Spanish group: Universitat Politécnica de Valencia; one Japanese group; and one American group: LCC.

In total, 15 runs have been returned by the participants: eight monolingual runs (FR-to-FR) and seven bilingual runs (6 EN-to-FR, 1 PT-to-FR).

It appears that the number of participants for the French task is the same that last year but it's the first time there are non-European participants. This shows there is a new major interest for the French as target language.

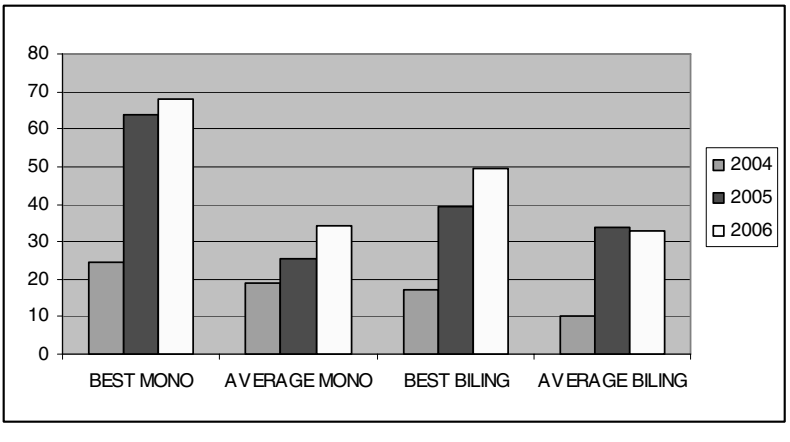

Fig. 3. Best and average scores for systems using French as target in CLEF QA campaigns

Two groups submitted four runs, two other groups submitted two runs and three groups submitted only one run.

Figure 3 shows the best and the average scores for systems using French as target in the last three CLEF QA campaigns.

For both monolingual and bilingual tasks, the best results were obtained by a French group, Synapse Développement. Another French group, LIA, reached the 2nd position for the two tasks.

Table 12 and 13 shows the results of the assessment of each run for each participant and for the two tasks.

The For the monolingual task, the systems returned between 27 and 129 correct answers in 1st rank.

For the bilingual task, the systems returned between 19 and 86 correct answers.

test set was composed of 190 Factual (F), Definition (D) and Temporally restricted (T) questions, and 10 List questions.

The accuracy has been calculated over all the first answers of F, D, T questions and also the Confidence Weighted Score (CWS), the Mean Reciprocal Rank score (MRR) and the K1 measure.

For the List questions, the P@N has been calculated.

For the monolingual task, the best system returned 67.89 % of correct answers (overall accuracy in 1st rank). We can observe this system obtained better results for definition questions (83.33 %) than for Factoid questions (63.51 %).

The LIA' system, which reached the second position in this task, returned 46.32 % of correct answers (overall accuracy in 1st rank). We can also observe the difference between the results for the Factual questions and the results for the Definition questions: 37.84 % of correct answers for the Factual and 76.19 % for the Definition questions.

Table 12a. Assessment of the monolingual and bilingual French

Id Participant	Assessed Answers (#)	Right answers (#)	Wrong answers (#)	ineXact answers (#)	U answers (#)
aske061frf	635	27	138	12	12
lcea061frf	589	30	151	6	3
lina061frfr	207	56	114	18	2
syna061fr	**200**	**129**	**50**	**9**	**2**
ulia061frfr	200	88	93	7	2
ulia062frfr	200	86	89	9	6
upv061frfr	200	60	119	10	1
upv062frfr	200	47	124	18	1
aske061en	640	19	157	6	8
lcc061enfr	578	40	125	23	2
syna061e	**200**	**86**	**97**	**6**	**1**
syna062en	200	63	120	6	1
ulia061enf	200	66	114	7	3
ulia062enf	200	66	111	9	4
syna061pt	**200**	**94**	**90**	**4**	**2**

For the bilingual task, the best system obtained 45.26 % of correct answers as opposed to 34.74 % of correct answers for the LIA' system.

We can remark that the best system for the bilingual task (EN-to-FR) obtained worse results than the second system for the monolingual task.

This year, before the assessment, the French assessors determined some rules to face up to problems encountered the last year.

Concerning Temporally restricted questions for example, to assess an answer as "Correct", the date, the period or the event had to be present in the document returned by the systems.

They decided also to check separately, at the end of the assessment, some questions which seemed difficult to them, to make sure that each answer had received the same "treatment" during the evaluation.

The main problem encountered this year, was related to the assessment of the List questions. This was a new kind of questions this year and participants followed different ways to answer to these questions. Some systems returned a list of answers in a same line; others returned an answer per line. ELDA evaluated these answers according to each run (if a line contained one of correct answers or all the correct answers, these answers had been assessed as "Correct".

The best system obtained 5 correct answers out of 10 List questions in total.

We can observe that the results for the List questions were not very relevant because of not much questions and not much rules.

Table 12b. Assessment of the monolingual and bilingual Frenchv

Id Participant	Overall Accuracy (%)	Accuracy over F (%)	Accuracy over D (%)	MRR (F, D, T)	CWS (F, D, T)	K1 Measure	P@N (L)
aske061frf	14.21	16.89	4.76	0.1974	0.1421	---	0.0900
lcea061frfr	15.79	10.14	35.71	0.1907	0.1578	---	0.1633
lina061frfr	29.47	27.70	35.71	0.2947	0.2551	-0.3777	0.3651
syna061fr	**67.89**	**63.51**	**83.33**	**0.6789**	**0.5568**	**0.2729**	**0.5000**
ulia061frfr	46.32	37.84	76.19	0.4632	0.4607	0.0684	0.5000
ulia062frfr	45.26	36.49	76.16	0.45016	0.4501	0.0474	0.2000
upv061frfr	31.58	31.08	33.33	0.3158	0.1638	-0.0047	0.3000
upv062frfr	24.74	26.35	19.05	0.2474	0.1088	-0.0931	0.2000
aske061en	10.00	12.16	2.38	0.1445		-0.2797	0.0633
lcc061enfr	21.05	25.00	7.14	0.2623		-0.1816	0.3967
syna061en	**45.26**	**37.16**	**73.81**	**0.4526**	**0.4526**	---	**0.2000**
syna062en	33.16	25.68	59.52	0.3316	0.3315	---	0.1000
ulia061enf	34.74	26.35	64.29	0.3474	0.334	-0.1789	0
ulia062enf	34.74	26.35	64.29	0.3474	0.3474	-0.1789	0.1000
syna061pt	**49.47**	**41.50**	**76.74**	**0.4947**	**0.4947**	**---**	**0**

In conclusion, this year, a system obtained "excellent" results. Synapse Développement obtained 129 correct answers out of 200 (as opposed to 128 last year).

This system is the best system for the French language. This year, it's again the dominant system.

In addition, we can observe the same great interest in Question Answering from the European (and now non-European) research community for the tasks using French as target language.

6.5 German as Target

Three research groups submitted runs for evaluation in the track having German as target language: The German Research Center for Artificial Intelligence (DFKI), FernUniversität Hagen (FUHA) and The Institute for Natural Language Processing in Stuttgart (IMS).

Table 13. Best and Aggregated Mono; Best and Aggregated Cross

Year	Best Mono	Aggregated Mono	Best Cross	Aggregated Cross
2006	42.33	64.02	32.98	33.86
2005	43.5	58.5	23	28
2004	34.01	43.65	0	0

All of them provided system runs for the monolingual scenario and just one group (DFKI) submitted runs for the cross-language English-German scenario.

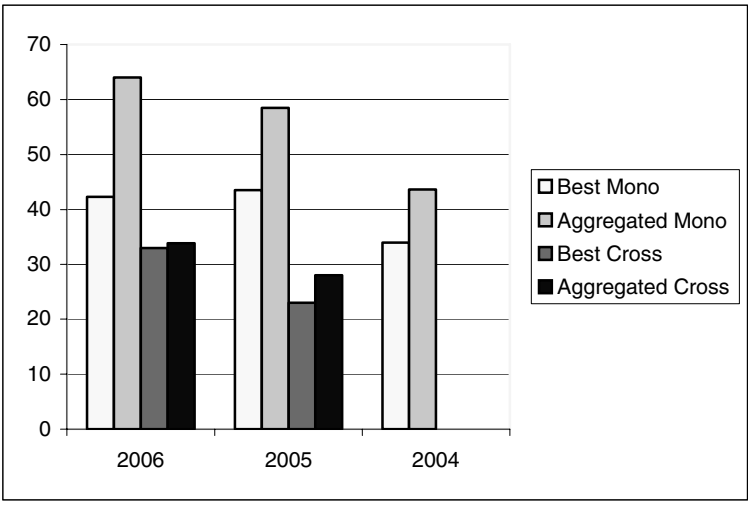

Fig. 4. Results evolution

Table 14. Performance of evaluated systems

Run ID	# Right			# ineXact			# Unsupported			Accuracy	MRR
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd		
dfki061dede$_M$	**80**	8	7	6	6	1	8	4	1	**42.32**	**45.67**
dfki062dede$_M$	63	15	3	4	5	3	8	0	2	33.33	37.83
fuha061dede$_M$	61	0	0	0	0	0	4	0	0	32.27	32.27
fuha062dede$_M$	64	0	0	1	0	0	4	0	0	33.86	33.86
ims061dede$_M$	25	2	3	0	1	0	8	2	0	13.22	14.28
ims062dede$_M$	23	3	2	0	1	0	8	2	0	12.16	13.31
dfki061ende$_C$	62	5	7	3	4	2	6	3	0	32.8	35.36
dfki062ende$_C$	50	10	2	5	4	2	3	2	1	26.45	29.45

Two assessors with different profiles conducted the evaluation: a native German speaker with little knowledge of QA systems and a researcher with advanced knowledge of QA systems and a good command of German. Compared to the previous editions of the evaluation forum, this year an increase in the performance of an aggregated virtual system for both monolingual and cross-language tasks was registered, as well as for the cross-language best system's result (Figure 4, Table 13). Given the increased complexity of the task (no question type provided, supporting snippets required) and of questions (factoid, definition and list), the stability of the best monolingual results can be considered also a gain in terms of performance.

Table 15a. System Performance: details

Run ID	Right		W	X	U	% F	% T	% D
	#	%	#	#	#	[152]	[44]	[37]
dfki061dede$_M$	80	42.32	95	6	8	38.81	29.54	56.75
dfki062dede$_M$	63	33.33	114	4	8	30.92	22.72	43.24
fuha061dede$_M$	61	32.27	124	0	4	30.26	15.9	40.54
fuha062dede$_M$	64	33.86	120	1	4	31.57	18.18	43.24
ims061dede$_M$	25	13.22	156	0	8	13.81	9	10.81
ims062dede$_M$	23	12.16	158	0	8	12.5	6.81	10.81
dfki061ende$_C$	62	32.8	117	3	6	28.94	22.72	48.64
dfki062ende$_C$	50	26.45	130	5	3	22.36	20.45	43.24

Concerning factoid questions, their increased complexity is reflected in systems being required to use some sort of lexical inference in order to track down the relevant contexts of the right answer, contexts that may be scattered across several adjoining sentences.

Table 15b. System Performance: Details

Run ID	% D	P@N L	NIL [20]			CWS	K1
	[37]	[9]	F	P	R		
dfki061dede$_M$	63.64	25.93	0.35	0.28	0.45	0	0
dfki062dede$_M$	48.48	33.33	0.32	0.27	0.4	0	0
fuha061dede$_M$	36.36	11.11	0.23	0.13	0.95	0.3	0.18
fuha062dede$_M$	39.39	11.11	0.24	0.14	0.95	0.32	0.19
ims061dede$_M$	9.09	25.42	0.2	0.12	0.55	0.07	-0.33
ims062dede$_M$	9.09	26.43	0.19	0.12	0.5	0.06	-0.33
dfki061ende$_C$	56.25	10	0.31	0.21	0.6	0	0
dfki062ende$_C$	50	10	0.33	0.22	0.65	0	0

Except for FUHA, the other two groups provided more than one possible answer per question, of which only the first three were manually evaluated. In order to come up with a measure of performance for systems providing several answers per question, Mean Reciprocal Rank (MRR) over right answers has been considered for this purpose.

Table 14 resumes the distribution of the right, inexact and unsupported answers over the first three ranked positions as delivered by the systems, as well as the accuracy and MRR for each of the runs.

Table 16. Inter-Assessor Agreement/Disagreement (breakdown)

Run ID	# Questions	# Answers	# Q-Disagreements				# A-Disagreements			
			Total	F	D	L	Total	X	U	W/ R
dfki061dede$_M$	198	437	35	28	7	0	44	20	16	8
dfki062dede$_M$	198	476	28	19	6	3	40	13	19	8
fuha061dede$_M$	198	198	12	8	4	0	11	3	2	6
fuha062dede$_M$	198	198	13	8	5	0	12	4	2	6
ims061dede$_M$	198	432	15	13	0	2	30	13	9	8
ims062dede$_M$	198	436	17	15	0	2	28	5	14	9
dfki061ende$_C$	198	405	26	20	5	1	33	12	16	5
dfki062ende$_C$	198	402	27	21	6	0	35	21	10	4

Two things can be concluded from the answer distribution of Table 14: first, there are a fair number of inexact and unsupported answers that show performance could be improved with a better answer extraction; second, the fair number of right answers among the second and third ranked positions indicate that there is still place for improvements with a more focused answer selection.

The details of systems' results can be seen in Table 15, in which the performance measures has been computed only for the first ranked answers to each question, except for the list questions. As an overall remark on the results, the majority of the evaluated system runs registered a better performance on the definition than on the factoid and temporal questions.

Table 16 describes the inter-rater disagreement on the assessment of answers in terms of question and answer disagreement. Question disagreement reflects the number of questions on which the assessors delivered different judgments and answer disagreement is a figure of the total number of answers disagreed on. Along the total figures for both types of disagreement, a breakdown at the question type level (Factoid, Definition, List) and at the assessment value level (ineXact, Unsupported, Wrong/Right) is listed.

The answer disagreements of type Wrong/Right are trivial errors during the assessment process when a right answers was considered wrong by mistake and the other way around, while those of type X or U reflect different judgments whereby an assessor considered an answer inexact or unsupported while the other marked it as right or wrong.

6.6 Italian as Target

Two groups participated in the Italian monolingual task, ITC-irst and the Universidad Politécnica de Valencia (UPV); while one group, the Università La Sapienza di Roma, participated in the cross-language EN-IT task. In total, five runs were submitted.

For the first time a cross-language task with Italian as target was chosen to test a participating system.

The best performance in the monolingual task was obtained by the UPV, which achieved an accuracy of 28.19%. Almost the same result was recorded last year (see Figure 5). The average accuracy in the monolingual task was 26.41%, which is an improvement of more than 2% with respect to last year's results.

The accuracy in the bilingual task was 17.02%, achieved by both submitted runs.

During the years the overall accuracy has steadily decreased starting from a 25.17% in the 2004, we reached a 24.08% in the 2005 and 22.06% this year. This could be partly due to newcomers – who usually get lower scores – and first experiments with bilingual tasks.

From the results shown in Table 17, it can be seen that the Universidad Politécnica de Valencia (UPV) submitted two runs in the monolingual task and achieved the best overall performance. The accuracy over Definition and Factoid questions ranged from 26.83% to 29.27%. ITC-Irst submitted one run, and achieved much better accuracy over Factoid questions (25.00%) than over Definition questions (17.07%).

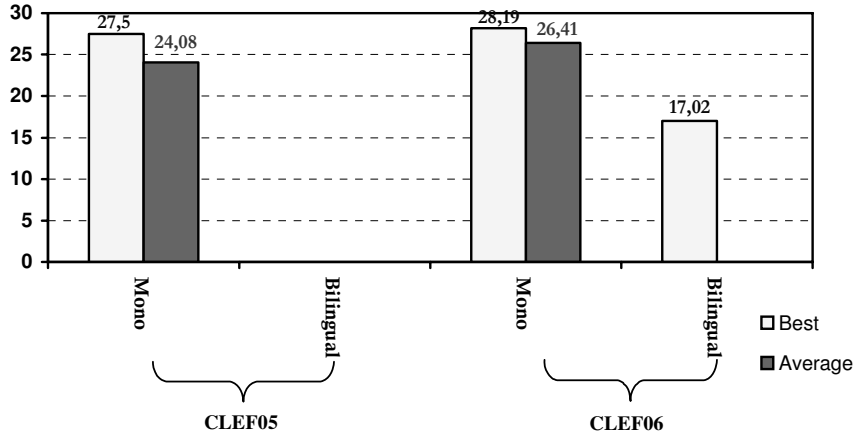

Fig. 5. Best and Average performance in the Monolingual and Bilingual tasks

As previously mentioned, the Università La Sapienza di Roma submitted two runs in the cross-language EN-IT tasks, performing much better in the Definition questions (24.39%)%) than in the Factoid questions (15.28%).

Table 17. Results of the monolingual and bilingual Italian runs

Run Name	Right answers (#)	wrong answers (#)	ineXact answers (#)	Unsupported answers (#)	Overall Accuracy (%)	Accuracy over F (%)	Accuracy over D (%)	P@N for L	Confidence Weighted Score
irst06itit	43	121	10	13	22.87	25.0	17.07	0.1528	0.19602
upv_061itit	53	124	6	5	28.19	28.47	26.83	0.0833	0.12330
upv_062itit	53	127	4	3	28.19	27.78	29.27	0.1667	0.13209
Roma061enit	32	141	4	11	17.02	15.28	24.39	0.1000	0.08433
Roma062enit	32	141	4	11	17.02	15.28	24.39	0.1500	0.08433

As far as List questions are concerned, all participating systems performed rather poorly, with a P@N ranging from 0.08 to 0.17. This implies that a more in-depth research on these questions and the measures for their evaluation is still needed.

Table 18. Temporally Restricted Questions: Right, Unsupported and Wrong Answer and Accuracy

	R	U	W	Accuracy %
irst06itit	6	6	26	15.79
Roma061enit	2	4	32	5.26
Roma062enit	2	4	32	5.26
upv_061itit	8	0	30	21.05
upv_062itit	9	0	29	23.68

Temporally restricted questions represented a challenge for the systems, which generally achieved a lower than average accuracy in this sub-category. The Universidad Politécnica de Valencia achieved the best performance of 23.68% (see Table 18). The evaluation process did not presented particular problems, although it was more demanding than usual because of the necessity to check the supporting text snippet. All runs were anyway assessed by two judges. The inter-assessor agreement was averagely 90,14 %, most disagreement being between U and X. A couple of cases of disagreement between R and W were due just to trivial mistakes.

6.7 Portuguese as Target

This year five research groups took part in tasks with Portuguese as target language, submitting ten runs: seven in the monolingual task, two with English as source, and one with Spanish. Two new groups joined for Portuguese: University of Porto, and Brazilian NILC, while LCC participated with an English-Portuguese run only. Universidade de Évora did not participate this year.

Table 19 presents the overall results concerning all questions. We present values both taking into account only the first answer to each question, and – for the only system where this makes any difference – all answers, assessing as right (or partially right) if any answer, irrespective of position, was right (or partially right). We have also distinguished inexact answers (X) between too little and too much information, respectively coded as X- and X+. There were only 18 NIL questions in the Portuguese collection.

Just like last year, Priberam achieved the best results by a clear margin. Also, their Spanish-Portuguese run, *prib061espt*, despite using a different (closely related) language as source, managed to achieve the second best result..

On the other hand, overall results for both Priberam and Esfinge show but a small improvement compared to 2005

It remains to be seen whether this year's questions displayed a higher difficulty or whether the systems themselves were subject to few changes.

We also provide in Table 20 the overall accuracy considering (and evaluating) independently all different answers provided by the systems.

Table 19. Results of the runs with Portuguese as target: first answers only, and all answers (marked with *)

Run Name	R (#)	W (#)	X+ (#)	X- (#)	U (#)	Overall Accuracy (%)	Accuracy over F (%)	Accuracy over D (%)	NIL Accuracy	
									Precision	Recall
esfg061ptpt	49	139	7	2	3	24.5	22.88	29.79	15.53	88.89
esfg062ptpt	45	142	6	6	1	22.5	20.26	29.79	14.95	88.89
nilc061ptpt	0	189	1	8	2	0.0	0.00	0.00	-	-
nilc062ptpt	3	190	0	5	2	1.5	1.96	0.00	8.57	16.67
prib061ptpt	134	58	6	1	1	67.0	65.36	72.34	43.33	72.22
uporto061ptpt	23	177	0	0	0	11.5	9.80	17.02	8.29	77.78
uporto062ptpt	26	169	2	3	0	13.0	11.76	17.02	7.64	66.67
esfg061enpt	27	164	3	2	2	13.5	11.76	19.15	11.35	88.89
lcc_061enpt	18	166	2	10	4	9.0	9.15	8.51	21.43	16.67
lcc_061enpt*	61	112	3	18	7	30.5	36.1	12.5	-	-
prib061espt	67	124	2	2	5	33.5	26.97	51.06	17.65	33.33

Table 20. Results of the runs with Portuguese as target: all answers

Run Name	R (#)	W (#)	X+ (#)	X- (#)	U (#)	Overall Accuracy %)
esfg061ptpt	49	143	11	2	3	23.56
esfg062ptpt	45	146	7	6	1	21.95
nilc061ptpt	0	189	1	8	2	0.00
nilc062ptpt	3	190	0	5	2	1.50
prib061ptpt	134	58	6	1	1	67.00
uporto061ptpt	36	178	0	0	0	16.82
uporto062ptpt	42	172	3	6	0	18.83
esfg061enpt	27	168	3	2	2	13.36
lcc_061enpt	141	1209	11	49	50	9.66
prib061espt	67	124	2	2	5	33.50

Table 21. Results of the assessment of the monolingual Portuguese runs: first answers only, except for lists, for which (for this table) one correct member of the list means the answer is to be considered correct.

	correct answers											
	Definition (47)				Factoid (t.r.q. ; list) (153 (27; 9))						Total	
Run	**Object** 7	**Organisation** 8	**Other**. 24	**Person** 9	**Location** 25 (2;0)	**Measure**21 (2;0)	**Organisation**23 (6;3)	**Other** 30 (6;2)	**Person** 34 (11;3)	**Time** 19 (0;1)	# 200 (27;9)	%
esfg061ptpt	3	3	5	3	10	4 (1;0)	2 (1;0)	6 (0;0)	11 (2;1)	3(0;0)	49 (4;1)	24.5
esfg062ptpt	3	4	5	2	9	4 (1:0)	1 (1; 0)	6 (1;0)	8 (2;1)	3 (0;0)	45 (4;1)	22.5
nilc061ptpt	0	0	0	0	0	0	0	0	0	0	0	0.0
nilc062ptpt	0	0	0	0	1	1 (1:0)	0	0 (1;0)	0	1(0;0)	3 (1;0)	1.5
prib061ptpt	**5**	**6**	**14**	**8**	**17 (1:0)**	**13 (1;0)**	**17 (3;3)**	**17 (1;1)**	**20 (4;3)**	**16 (0;1)**	**134 (10;8)**	**67.0**
uporto061ptpt	1	0	2	5	3	3 (1;0)	1 (1;0)	2 (0;0)	5 (2;1)	1(0;0)	23 (4;1)	11.5
uporto062ptpt	1	0	2	5	4	3 (1;0)	1 (1;0)	3 (0;0)	4 (2;1)	3(0;0)	26 (4;1)	13.0
combination	5	7	19	9	19 (1;0)	14 (1;0)	17 (3;3)	19 (1;1)	23 (5;3)	17 (0;1)	149 (11;8)	74.5
esfg061enpt	**1**	**2**	**2**	**4**	**7**	**3 (1:0)**	1 (1;1)	**4 (0;0)**	**2 (2;1)**	1 (0;0)	27 (4;2)	15.4
lcc_061enpt	**1**	1	**2**	0	2	1	**2 (1; 0)**	1 (0;1)	**2 (1;0)**	**6** (0;0)	18 (2;1)	9.0
pribe061espt	2	4	10	8	7	9 (1;0)	7 (0; 0)	6 (1;0)	7 (2;0)	7 (0;0)	67 (3;0)	33.5

Lcc's performance slightly increases , from 9.0 to 9.66, but if we are interested in correct answers anywhere, then Lcc* got the second best accuracy, 30.5, which might suggest that they might significantly improve their score by reranking mechanisms.

Table 21 shows the results for each answer type. In parentheses we display the subset of temporally-restricted questions, and we add the list questions, in order to provide the full picture.

Table 22. Size of answers and justifying snippets, in words

Run name	Answers (#)	Non-NIL Answers (#)	Average answer size	Average answer size (R only)	Average snippet size	Average snippet size (R only)
esfg061ptpt	208	105	3.4	3.2	108.8	108.5
esfg062ptpt	204	98	3.8	3.5	109.1	105.5
nilc061ptpt	200	200	5.7	-	5.7	-
nilc062ptpt	200	165	4.9	-	4.4	-
prib061ptpt	200	170	3.7	3.8	31.5	30.3
uporto061ptpt	210	29	3.1	3.2	39.7	32.7
uporto062ptpt	216	59	3.0	2.8	43.1	33.7
esfg061enpt	202	61	3.5	3.5	95.3	106.1
lcc_061enpt	1463	1449	5.2	4.1	35.2	34.6
prib061espt	200	166	3.5	4.3	31.3	29.1

A virtual run, called combination, was included in Table 21 and computed as follows: if any of the participating systems found a right answer, it is considered right in the combination run. Ideally, this combination run measures the potential achievement of cooperation among all participants. However, for Portuguese this combination does not significantly outperform the best performance: Priberam alone corresponds to 89.9% of the combination run.

Table 23. Accuracy of temporally restricted questions (all answers considered), compared to non-temporally restricted ones, and to overall accuracy

Run name	Questions with at least one correct answer (#)	Accuracy for T.R.Q. (%)	Accuracy for non-T.R.Q (%)	Total accuracy (%)
esfg061ptpt	3	11.11	25.41	23.56
esfg062ptpt	3	11.11	23.60	21.95
nilc061ptpt	0	0.00	0.00	0.00
nilc062ptpt	1	2.86	1.16	1.50
prib061ptpt	10	37.04	71.68	67.00
uporto061ptpt	4	14.81	17.11	16.82
uporto062ptpt	4	14.81	19.39	18.83
esfg061enpt	3	11.11	13.71	13.37
lcc_061enpt	7	2-86	11.03	9.66
prib061espt	3	11.11	36.99	33.50

Table 24. Results for List Questions

Question	Known answers	Esfig061ptpt	esfig062ptpt	nilc061ptpt	nilc062ptpt	pribo61ptpt	uporto061ptpt	uporto062ptpt	esfig061enpt	lcc061enpt	esfig061espt
205	3	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	0/10	0/1
399	3	0/1	0/1	0/1	0/1	3/3	0/1	0/1	0/1	3/9	0/1
400	3	0.5/3	0.5/3	0/1	0/1	3/3	0/1	0/1	0.5/3	0/8	3/3
759	3	0/1	0/1	0.5/1	0.5/1	1/1	0/1	0/1	0/1	0.5/10	0/1
770	3	0.5/1	0.5/1	0/1	0/1	1/1	0/1	0/1	0/1	2/10	0/1
784	5	0/1	0/1	0/1	0/1	1/1	0/1	0/1	0/1	1/9	0/1
785	3	0/1	0/1	0/1	0/1	0.5/1	0/1	0/1	0/1	0/10	0/1
786	3	0/1	0/1	0.5/1	0.5/1	1/1	0/1	0/1	0/1	2/10	0/1
795	5	0/1	0/1	0/1	0/1	1/1	0/1	0/1	0/1	2/7	0/1
score		0.030	0.030	0.037	0.037	0.396	0	0	0.019	0.113	0.011

We have also analysed the size in words of both answers and justification snippets, as displayed in Table 22. (Computations were made excluding NIL answers.) Interestingly, Priberam provided the shortest justifications.

In Table 23, we compare the accuracy of the systems for the 22 temporally restricted questions in the Portuguese question set with their scores for non-temporally restricted ones and their overall performance.

Finally, a total of nine questions were defined by the organization as requiring a list as proper answer. The fact that the systems had to find out whether multiple or single answers were expected was a new feature this year and was not conveniently handled by most systems.

In fact, two systems (Priberam and NILC) completely ignored this and provided a single answer to every question, while two other systems, although attempting to deal with list questions, seemed to fail in appropriately identifying them: RAPOSA (UPorto) provided multiple answers only to non-list questions, and Esfinge produced 11 answers for the nine questions.

In fact, only LCC presented multiple answers systematically, yielding an average of 7.32 answers per question, while no other group exceeded 1.1. For the case of closed lists (where "one" answer might bring all answers, such as "Lituânia, Estónia e Letónia"), we still counted the number of answers individually (3).

We believe further study should be devoted to the list questions for the next years, since a distinction between closed lists and open lists, although acknowledged, was not properly taken into consideration. We have thus chosen to handle all these questions alike, assigning them the following accuracy score: number of correct answers (where X counted as ½), presented left of the slash, divided by the sum of the number of existing answers in the collections and the number of wrong distinct answers provided by the system, right of the slash. The results are displayed in Table 24.

6.8 Spanish as Target

The participation at the Spanish as Target subtask is still growing. Nine groups, two more than the last year, submitted 17 runs: 12 monolingual, 3 from English, 1from French and 1 from Portuguese. Table 25 and Table 26 show the summary of systems results for monolingual and cross-lingual respectively. The number of Right, Wrong (W), Inexact (X) and Unsupported (U) answers

Table 25a. Results at the Spanish as target, monolingual

Run	Right		W	X	U	% F	% T	% D	% L
	#	%	#	#	#	[108]	[40]	[42]	[10]
pribe061	105	**52,50**	86	4	5	**55,56**	30,00	69,05	**40,00**
inao061	102	51,00	86	3	9	47,22	**35,00**	**83,33**	20,00
vein061	80	40,00	112	3	5	32,41	25,00	**83,33**	-
alia061	72	36,00	105	15	8	38,89	22,50	50,00	-
upv_061	70	35,00	119	5	6	37,04	25,00	47,62	-
upv_062	57	28,50	123	6	14	27,78	25,00	40,48	-
aliv061	56	28,00	123	8	13	29,63	22,50	35,71	-
aliv062	56	28,00	132	6	6	26,85	25,00	40,48	-
mira062	41	20,50	148	4	7	21,30	17,50	23,81	10,00
sinaiBruja06	39	19,50	146	6	9	16,67	17,50	33,33	-
mira061	37	18,50	154	3	6	21,30	15,00	16,67	10,00
aske061	27	13,50	143	1	29	15,74	12,50	9,52	10,00

Table 25b. Results at the Spanish as target, monolingual

Run	NIL [20]			r	% Answer Extraction
	F	P	R		
pribe061	0,44	0,34	0,60	-	84,68
inao061	**0,46**	0,38	0,60	0,216	**86,44**
vein061	0,34	0,21	0,80	0,133	86,02
alia061	0,34	0,22	0,75	**0,322**	69,23
upv_061	0,43	0,33	0,65	0,194	70,71
upv_062	0,41	0,32	0,60	0,163	66,28
aliv061	0,34	0,33	0,35	0,190	65,12
aliv062	0,33	0,26	0,45	0,153	72,73
mira062	0,35	0,35	0,35	0,145	43,62
sinaiBruja06	0,23	0,13	0,90	-0,119	79,59
mira061	0,34	0,26	0,50	0,136	51,39
aske061	0,08	0,20	0,05	0,199	62,79

Tables show also the accuracy (in percentage) of factoids (F), factoids with temporal restriction (T), definitions (D) and list questions (L). Best values are marked in bold face. Best performing systems have improved their performance (as seen in Figure 5), mainly with respect to factoids

However, performance when the question has a temporal restriction didn't vary significantly. Last year, the answering of definitions with respect to persons and organizations was almost solved. In spite of the fact that this year the set of definition questions was more realistic systems have improved slightly their performance.

List questions have been introduced this year so they deserve some attention regarding their evaluation. We have differentiated two types of list questions: conjunctive and disjunctive (as presented in [4]).

Conjunctive list questions are asking for a set of items and they are *Right* if all the items are present in the answer. For example, *"Nombre los tres Beatles que siguen vivos"* (Name the three Beatles alive).Disjunctive list questions are asking for an undetermined number of items. For example, *"Nombre luchadores de Sumo"* (Name Sumo fighters). Only the first answer of each system has been evaluated in both cases.

Table 26a. Results at the Spanish as target, Cross-lingual

Run	Right		W	X	U
	#	%	#	#	#
pribe061ptes	72	36,00	123	3	2
alia061enes	41	20,50	134	9	16
lcc_061enes	38	19,00	141	14	7
aske061fres	23	11,50	162	-	15
aske061enes	12	6,00	178	-	10

Table 26b. Results at the Spanish as target, Cross-lingual

Run	% F	% T	% D	% L	NIL [20]				% Answer Extraction
	[108]	[40]	[42]	[10]	F	P	R	r	
pribe061ptes	39,81	27,50	38,10	20	0,29	0,29	0,30	-	78,26
alia061enes	17,59	12,50	40,48	-	0,31	0,19	0,80	0,142	65,08
lcc_061enes	20,37	25,00	14,29	-	0,35	0,35	0,35	0,067	55,07
aske061fres	13,89	10,00	7,14	10	0,08	0,17	0,05	0,302	53,49
aske061enes	6,48	2,50	7,14	10	0,10	1,00	0,05	0,091	40,00

Regarding the NIL questions, Table 25 and 26 show the harmonic mean (F) of precision (P) and recall (R). The best performing systems have increased again their performance (see Table 27) in NIL questions. The correlation efficient r between the self-score and the correctness of the answers has been increased in the majority of systems, although results are not good enough yet.

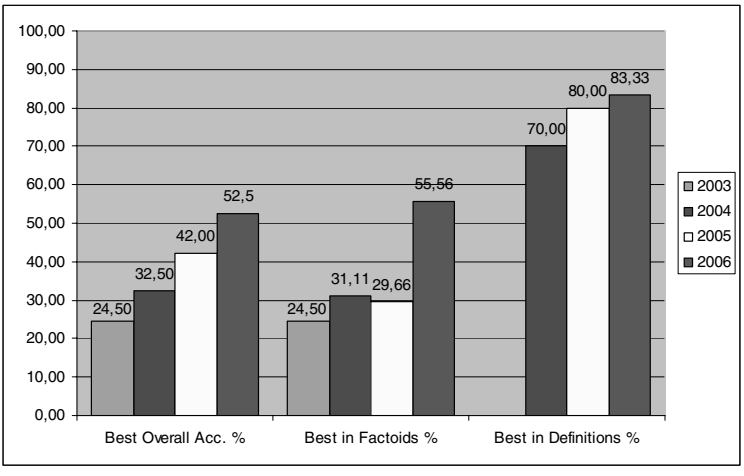

Fig. 6. Evolution of best performing systems 2003-2006

This year a supporting text snippet was requested. For this reason, we have evaluated the systems capability to extract the answer when the snippet contains it.

The last column of Tables 25 and 26 shows the percentage of cases where the correct answer was correctly extracted. This information is very useful to diagnose if the lack of performance is due to the passage retrieval or to the answer extraction.

Regarding Cross-Lingual runs, it is worth to mention that Priberam has achieved in the Portuguese to Spanish task a result comparable to the monolingual runs.

Table 27. Evolution of best results in NIL questions

Year	F-measure
2003	0,25
2004	0,30
2005	0,38
2006	0,46

All the answers have been assessed anonymously considering all systems' answers simultaneously question by question. The inter-annotator agreement was evaluated over 985 answers assessed by the two judges. Only a 2.5% of the judgements were different and the resulting kappa value was 0.93.

7 Conclusions

The QA track at CLEF 2006 has once again demonstrated the interest for Question Answering in languages other than English. In fact, both the number of participants and runs submitted has grown, following the positive trends of the previous campaign.

The balance between tradition and innovations –i.e the introduction of list questions and supporting text snippets- has proved to be a good solution, which allows both new-comers and veterans to test their systems against adequately challenging tasks and, at the same time, to make a comparison with previous exercises. Generally speaking, the results recorded an improvement in performance, with best accuracy significantly higher than in previous campaigns both in monolingual and bilingual tasks.

As far as the organisation of the campaign is concerned, the introduction of new elements such as list questions and supporting snippets has implied a significant increase of work both in the question collection and in the evaluation phase, which was particularly demanding for language groups which had a great number of participants. A better distribution of the workload and solutions to speed up the evaluation process, also with automatic assessment of part of the submissions will be essential in next campaigns.

A future perspective of QA is certainly outlined by the two pilot tasks offered in 2006-i.e. AVE and WiQa-, the latter in particular representing a significant step toward a more realistic scenario, where queries are carried out on the Web. For these reasons, a quick integration of these experiments into the main task is hoped for.

Acknowledgements

The authors would like to thank Donna Harman for her valuable feedback and advice, and Diana Santos for her precious contribution in the organization of the campaign and the revision of this paper.

Paulo Rocha is thankful to the many useful comments and overall discussion with Diana Santos for the Portuguese part. He was also supported by the Portuguese Fundação para a Ciência e Tecnologia within the Linguateca project, through grant POSI/PLP/43931/2001, co-financed by POSI.

Bogdan Sacaleanu was supported by the German Federal Ministry of Education and Research (BMBF) through the projects HyLaP and COLLATE II.

Anselmo Peñas has been partially supported by the Spanish Ministry of Science and Technology within the project R2D2-SyEMBRA (TIC-2003-07158-C04-02).

References

1. QA@CLEF 2006 Organizing Committee. Guidelines (2006), http://clef-qa.itc.it/guidelines.html
2. WiQA Website, http://ilps.science.uva.nl/WiQA/
3. AVE Website, http://nlp.uned.es/QA/AVE/

4. Herrera, J., Peñas, A., Verdejo, F.: Question answering pilot task at CLEF 2004. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 581–590. Springer, Heidelberg (2005)
5. Magnini, B., Vallin, A., Ayache, C., Erbach, G., Peñas, A., de Rijke, M., Rocha, P., Simov, K., Sutcliffe, R.: Overview of the CLEF 2004 Multilingual Question Answering Track. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 371–391. Springer, Heidelberg (2005)
6. Rocha, P., Santos, D.: Abrindo a porta á participa internacional em avaliação de RI do português. In: Santos, D. (ed.) Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa, IST Press, Lisbon (in Press, 2006)
7. Santos, D., Rocha, P.: The Key to the First CLEF with Portuguese: Topics, Questions and Answers in CHAVE. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 821–832. Springer, Heidelberg (2005)
8. Spark Jones, K.: Is question answering a rational task? In: Bernardi, R., Moortgat, M. (eds.) Questions and Answers: Theoretical and Applied Perspectives. Second CoLogNETElsNET Symposium, pp. 24–35. Amsterdam (2003)
9. Vallin, A., Magnini, B., Giampiccolo, D., Aunimo, L., Ayache, C., Osenova, P., Peñas, A., de Rijke, M., Sacaleanu, B., Santos, D., Sutcliffe, R.: Overview of the CLEF 2005 Multilingual Question Answering Track. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 21–23. Springer, Heidelberg (2006)
10. Voorhees, E. M.: Overview of the TREC 2002 Question Answering Track. In: Voorhees, E.M., Buckland, L.P. (eds) Proceedings of the Eleventh Text Retrieval Conference (TREC 2002 NIST Special Publication 500-251, pp. 115–123. Washington DC (2002)