

OVERVIEW OF THE CLEF 2007 MULTILINGUAL QUESTION ANSWERING TRACK

Danilo Giampiccolo¹, Pamela Forner¹, Anselmo Peñas², Christelle Ayache³, Dan Cristea⁴,
Valentin Jijkoun⁵, Petya Osenova⁶, Paulo Rocha⁷, Bogdan Sacaleanu⁸, and Richard Sutcliffe⁹

¹ CELCT, Trento, Italy ({giampiccolo, forner}@celct.it)

² Departamento de Lenguajes y Sistemas Informáticos, UNED, Madrid, Spain (anselmo@lsi.uned.es)

³ ELDA/ELRA, Paris, France (ayache@elda.fr)

⁴ Faculty of Computer Science, University “Al. I. Cuza” of Iași, Romania Institute for Computer Science,
Romanian Academy, Iași, Romania (dcristea@info.uaic.ro)

⁵ Informatics Institute, University of Amsterdam, The Netherlands (jijkoun@science.uva.nl)

⁶ BTB, Bulgaria, (petya@bultreebank.org)

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.

The fifth QA campaign at CLEF, the first having been held in 2006. was characterized by continuity with the past and at the same time by innovation. In fact, topics were introduced, under which a number of Question-Answer pairs could be grouped in clusters, containing also co-references between them. Moreover, the systems were given the possibility to search for answers in Wikipedia. In addition to the main task, two other tasks were offered, namely the Answer Validation Exercise (AVE), which continued last year's successful pilot, and QUASt, aimed at evaluating the task of Question Answering in Speech Transcription. The Question Answering Real Time was also proposed again, after the success obtained in 2006.

As general remark, it must be said that the task proved to be more difficult than expected, as in comparison with last year's results the Best Overall Accuracy dropped from 49% to 41,75% in the multi-lingual subtasks, and, more significantly, from 68% to 54% in the monolingual subtasks..

1 Introduction

The fifth QA campaign at CLEF [1], the first having been held in 2003, was characterized by continuity with the past, maintaining the focus on cross-linguality and covering as many European languages as possible (with the addition of Indonesian); and by innovation 1) by introducing a number of Question-Answer pairs, grouped in clusters, which referred to a same topic and which contained co-references between them, and 2) by giving the possibility to search for answers in Wikipedia. In this way, the newcomers had the possibility to test themselves with the classic task, and those who had participated in the previous campaigns had a new challenging factor to test their systems. In addition to the main task, three other tasks were offered, namely the Answer Validation Exercise (AVE), which continued last year's successful pilot, Question Answering for Speech Transcripts (QAST), aimed at evaluating the task of Question Answering in Speech Transcription, and the Question Answering Real Time, which made its successful debut in Alicante, in 2006.

In the following sections, the main task and its preparation will be described. A presentation of the participants and the runs submitted will be also given, together with a description of the evaluation method and the results achieved.

2 Tasks

Following the procedure consolidated in previous years, in the 2007 campaign several different tasks were proposed:

1. a *main task*, divided into several monolingual and bi-lingual sub-tasks;
2. the *Answer Validation Exercise (AVE)*, which continued the successful experiment proposed in 2006. Systems were required to emulate human assessment of QA responses and decide whether an *Answer* to

- a *Question* is correct or not according to a given *Text*. Participating systems were given a set of triplets (Question, Answer, Supporting Text) and they had to return a boolean value for each triplet. Results were evaluated against the QA human assessments [1];
3. the *QA Answering on Speech Transcripts (QAST)*, a pilot task which aimed at providing a framework in which factual. Relevant points of this pilot were:
 - a. Comparing the performances of the systems dealing with both types of transcriptions.
 - b. Measuring the loss of each system due to the state of the art ASR technology.
 - c. In general, motivating and driving the design of novel and robust factual QA architectures for automatic speech transcriptions [2].
 4. *Question Answering in Real-Time (QART)*, aimed at evaluating the ability of QA systems to answer within a time constraint. This year two tasks were planned: one exercise via web and another exercise to be carried out in the Budapest workshop [3].

The AVE and QAST tasks are described in details in dedicated papers in this Working Notes.

As far as the main task is concerned, the consolidated procedure was followed, although some relevant innovations were introduced.

The systems were given a set of 200 questions -which could concern facts or events (F-actoid questions), definitions of people, things or organisations (D-efinition questions), or lists of people, objects or data (L-ist questions)- and were asked to return one exact answer, where *exact* meant that neither more nor less than the information required was given. Following the example of TREC, this year the exercise consisted of topic-related questions, i.e. clusters of questions which were related to the same topic and possibly contained co-references between one question and the others. Neither the question types (F, D, L) or the topics were given to the participants.

The answer needed to be supported by the docid of the document in which the exact answer was found, and by portion(s) of text, which provided enough context to support the correctness of the exact answer. Supporting texts could be taken from different sections of the relevant documents, and had to sum up to a maximum of 700 bytes. 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*. As in previous years, the exact answer could be exactly copied and pasted from the document, even if it was grammatically incorrect (e.g.: inflectional case did not match the one required by the question). Anyway, this year systems were also allowed to use NL generation in order to correct morpho-syntactical inconsistencies (e.g., in German, changing "dem Presidenten" into "der President" if the question implies that the answer is in Nominative case), and to introduce grammatical and lexical changes (e.g., QUESTION: *What nationality is X?* TEXT: *X is from the Netherlands => EXACT ANSWER: Dutch*).

Table 1: Tasks activated in 2007 (in green)

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

The subtasks 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.

Ten source languages were considered, namely, Bulgarian, Dutch, English, French, German, Indonesian, Italian, Portuguese, Romanian and Spanish. All these languages were also considered as target languages, except for Indonesian, which had no news collections available for the queries and, as was done in the previous campaigns, used the English question set translated into Indonesian (IN).

As shown in Table 1, 37 tasks were proposed:

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

Anyway, as Table 2 shows, not all the proposed tasks were then carried out by the participants.

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

	MONOLINGUAL	CROSS-LINGUAL
CLEF 2004	6	13
CLEF 2005	8	15
CLEF 2006	7	17
CLEF 2007	7	11

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. English was still both source and target language in the cross-language tasks.

As the format is concerned, this year both input and output files were formatted as an XML file (for more details see [4]).

3 Test Set Preparation

The procedure followed to prepare the test set was much different from that used in the previous campaigns. First at all, each organizing group, responsible for a target language, freely chose a number of topics. For each topic, one to four questions were generated. Topics could be not only named entities or events, but also other categories such as objects, natural phenomena, etc. (e.g. George W. Bush; Olympic Games; notebooks; hurricanes; etc.). The set of ordered questions were related to the topic as follows:

- The topic was named either in the first question or in the first answer
- The following questions can contain co-references to the topic expressed in the first question/answer pair.

Topics were not given in the test set, but could be inferred from the first question/answer pair. For example, if the topic was *George W. Bush*, the cluster of questions related to it could have been:

- Q1: *Who is George W. Bush?*
- Q2: *When was he born?*

Q3: *Who is his wife?*

The Table 3: Document collections used in CLEF 2007.

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 questions in the set were numbered from 1 to 200, with no indication about whether they were part of a cluster belonging to the same topic.

Another major innovation of this year's campaign concerned the corpora at which the questions were aimed at. In fact, beside the data collections composed of news articles provided by ELRA/ELDA, also Wikipedia was considered, capitalizing on the experience of the WiQA pilot task proposed in 2006. The Wikipedia pages in the target languages, as found in the version of the Wikipedia of November, 2006 could be used. XML and the HTML versions were available for download, even though any other versions of the Wikipedia files could be used as long as they dated back to the end of November / beginning of December 2006. All the answers to the questions had to be taken from "actual entries" or articles of Wikipedia pages - the ones whose filenames normally correspond to the topic of the article. Other types of data ("image", "discussion", "category", "template", "revision histories", any files with user information, and any "meta-information" pages), had to be excluded.

As far as the question types are concerned, as in previous years of QA@CLEF, the three following categories were still considered:

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

We consider the following 8 answer types for factoids:

- PERSON, e.g. Q: *Who was called the "Iron-Chancellor"?*

- TIME, e.g. A: *Otto von Bismarck.*
Q: *What year was Martin Luther King murdered?*
A: *1968.*
- LOCATION, e.g. Q: *Which town was Wolfgang Amadeus Mozart born in?*
A: *Salzburg.*
- ORGANIZATION, e.g. Q: *What party does Tony Blair belong to?*
A: *Labour Party.*
- MEASURE, e.g. Q: *How high is Kanchenjunga?*
A: *8598m.*
- COUNT, e.g. Q: *How many people died during the Terror of Pol Pot?*
A: *1 million.*
- OBJECT, e.g. Q: *What does magma consist of?*
A: *Molten rock.*
- OTHER, i.e. everything that does not fit into the other categories above.
Q: *Which treaty was signed in 1979?*
A: *Israel-Egyptian peace treaty.*

b) *Definition questions*, questions such as "What/Who is X?", and are divided into the following subtypes:

- PERSON, i.e. questions asking for the role/job/important information about someone,
Q: *Who is Robert Altmann?*
A: *Film maker.*
- ORGANIZATION, i.e. questions asking for the mission/full name/important information about an organization, e.g.
Q: *What is the Knesset?*
A: *Parliament of Israel.*
- OBJECT, i.e. questions asking for the description/function of objects, e.g.
Q: *What is Atlantis?*
A: *Space Shuttle.*
- OTHER, i.e. question asking for the description of natural phenomena, technologies, legal procedures etc., e.g.
Q: *What is Eurovision?*
A: *Song contest.*

c) *closed list questions*: i.e. questions that require one answer containing a determined number of items, e.g:

Q: *Name all the airports in London, England.*
A: *Gatwick, Stansted, Heathrow, Luton and City.*

As only one answer was allowed, all the items had to be presented in sequence, one next to the other, in one document of the target collections.

Table 4: Test set breakdown according to question type

	F	D	L	T	NIL
BG	158	32	10	12	0
DE	164	28	8	27	0
EN	161	30	9	3	0
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
RO	160	30	10	52	7

Besides, all types of questions could contain a temporal restriction, i.e. a temporal specification that provided important information for the retrieval of the correct answer, for example:

Q: *Who was the Chancellor of Germany from 1974 to 1982?*
 A: *Helmut Schmidt.*
 Q: *Which book was published by George Orwell in 1945?*
 A: *Animal Farm.*
 Q: *Which organization did Shimon Perez chair after Isaac Rabin's death?*
 A: *Labour Party Central Committee.*

Some questions could have no answer in the document collection, and in that case the exact answer was "NIL" and the answer and support docid fields were left empty. A question is assumed to have no right answer when neither human assessors nor participating systems could find one.

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

Each of the question sets was finally then translated into English, so that each group could translate another set into their own language, when preparing the cross-lingual data sets which had been activated.

4 Participants

After years of constant growth, the number of participants has decreased in 2007 [see Table 5]..

Table 5: Number of participating groups

	America	Europe	Asia	Australia	TOTAL
CLEF 2003	3	5	-	-	8
CLEF 2004	1	17	-	-	18
CLEF 2005	1	22	1	-	24
CLEF 2006	4	24	2	-	30
CLEF 2007	3	17	1	1	22

The geographical distribution has anyway remained almost the same, recording a new entry of a group from Australia. No participants took part to any Bulgarian tasks.

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
CLEF 2007	22	23	14

Also the number of submitted runs has decreased sensibly, from a total of 77 registered last year to 22 (see table 6). As in previous campaigns, a larger number of people chose to participate in the monolingual tasks, which once again demonstrated to be more approachable.

5 Evaluation

No changes were made as far the evaluation process is concerned- Human judges assessed the exact answer (i.e. the shortest string of words which is supposed to provide the exact amount of information to answer the question) 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 judgment of all the runs. As regards the evaluation measures the following measures:

- accuracy, as the main evaluation score, defined as the average of SCORE(q) over all 200 questions q;
- the K1 measure[6]:

$$K1(sys) = \frac{\sum_{r \in answers(sys)} score(r) \bullet eval(r)}{\#questions}$$

$$K1(sys) \in \mathbb{R} \wedge K1(sys) \in [-1,1]$$

where:

score (r) is the confidence score assigned by the system to the answer r and eval(r) depends on the judgment given by the human assessor.

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

K1(sys) = 0 is established as a baseline.

- the Confident 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 [2].

6 Results

As far as accuracy is concerned, scores were generally far lower this year than usual, as Figure 1 shows. In detail, Best accuracy in the monolingual task decreased by almost 15 points, passing from last year's 68.95% to 54%, while Best accuracy in cross-language tasks passed from 49.47% to 41.75% recording.

As far as average performances are concerned, this year a neat decrease has been recorded in the biligual tasks, which went from 22.8% to 10.9%. This was due also due to the presence of systems which participated for the first time, achieving very low score in tasks which are quite difficult also for veterans.

As a general remark, it can be said that the new factors introduced this year appear to have had an impact on the performances of the systems. As more than one participant has noticed, there has been not enough time to adjust the systems to the new requirements.

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.

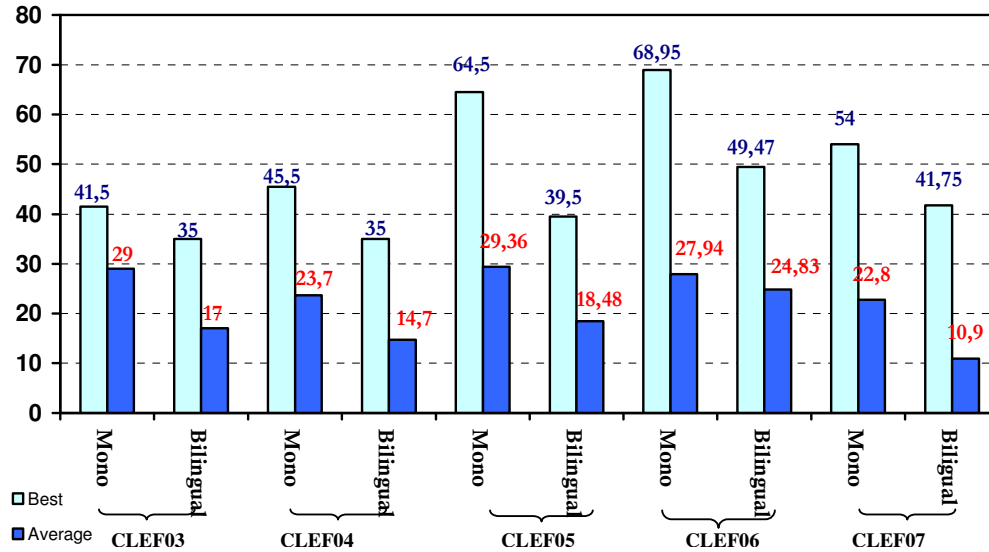


Figure 1: Best and average scores in CLEF QA campaigns

6.1 Dutch as Target

For the Dutch subtask of the CLEF 2007 QA task, three annotators generated 200 questions organized in 78 groups so that there were 16 groups with one question, 21 groups with two, 22 with three and 19 groups with four questions. Among the 200 questions 156 were factoids, 28 definitions and 16 list questions. In total, 41 questions had temporal restrictions. Table XXX below shows the distributions of topic types for groups and expected answer types for questions.

Table 6: Distribution of topic types and expected answer for questions.

Topic type	Number of topics
OBJECT	29
PERSON	18
ORGANIZATION	12
LOCATION	10
EVENT	19

Expected answer type	Number of questions
OTHER	45
PERSON	38
TIME	32
OBJECT	25
LOCATION	25
COUNT	14
ORGANIZATION	13
MEASURE	8

Annotators were asked to create questions with answers either in Dutch Wikipedia or in the Dutch newspaper corpus, as well as questions without known answers. Of 200 questions, 186 had answers in Wikipedia, and 14 in the newspaper corpus. Annotators did not create NIL questions.

Table 7: Results

Run	R #	W #	X #	U #	% F [156]	% T [41]	% D [28]	% L [16]	NIL		CWS	Overall accuracy
									#	% [0]		
uams071qrz	15	160	1	23	9.0	4.9	3.6	0	0	0	0.02	7.54
gron071NLNL	49	136	11	4	24.4	19.5	35.7	6.3	20	0	0.06	24.5
gron072NLNL	51	135	10	4	25.6	19.5	35.7	6.3	20	0	0.07	25.5
gron071ENNL	26	159	8	7	10.3	14.6	32.1	6.3	20	0	0.02	13
gron072ENNL	27	161	7	5	10.9	14.6	32.1	6.3	16	0	0.02	13.5

This year, two teams took part in the QA track with Dutch as the target language: the University of Amsterdam and the University of Groningen. The latter submitted both monolingual and crosslingual (English to Dutch) runs. The 5 submitted runs were assessed independently by 3 Dutch native speakers in such a way that each question group was assessed by at least two assessors. In case of conflicting assessments, assessors were asked to discuss the judgements and come to an agreement. Most of the occurred conflicts were due to difficulties in distinguishing between *inexact* and *correct* answers. Table 7 below shows the evaluation results for the five submitted runs (three monolingual and two cross-lingual). The table shows the number of Right, Wrong, inexact and Unsupported answers, as well as the percentage of correctly answered Factoids, Temporally restricted questions, Definition and List questions.

The best monolingual run (gron072NLNL) achieved accuracy of 25.5%, which is slightly less than the best results in the 2006 edition of the QA task. The same tendency holds for the performance on factoid and definition questions. We interpret this as an indication of the increased difficulty of the task due to the newly introduced Wikipedia collection.

One of the runs contained as many as 23 unsupported answers—this might indicate a bug in the system.

6.1 English as Target

Creation of questions. This year the questions set were radically different from last year. Instead of 200 independent questions, we were required to devise questions in groups. Each group had a declared topic (e.g. "Polygraph") but unlike in TREC, this topic was not communicated to the participants. As at CLEF last year, the type of question (e.g. definition, factoid or list) was not declared to participants either.

Table 8: Results

Run	R #	W #	X #	U #	% F [161]	% T [3]	% D [30]	% L [9]	NIL		CWS	KI	Overall accuracy
									#	% [0]			
cind071fren	26	171	1	2	11.18	0.00	23.33	11.11	0	0.00	0.00	0.00	13.00
cind072fren	26	170	2	2	11.18	0.00	23.33	11.11	0	0.00	0.00	0.00	13.00
csui071inen	20	175	4	1	10.56	0.00	10.00	0.00	0	0.00	0.00	0.00	10.00
dfki071deen	14	178	6	2	4.35	0.00	23.33	0.00	0	0.00	0.00	0.00	7.00
dfki071esen	5	189	4	2	1.86	0.00	6.67	0.00	0	0.00	0.00	0.00	2:50
mqa071nlen	0	200	0	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00
mqa072nlen	0	200	0	0	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00
wolv071roen	28	166	2	4	9.32	0.00	43.33	0.00	0	0.00	0.00	0.00	14.00

160 Factoids (in groups) were requested, together with 30 definitions and ten lists. The numbers of temporally restricted factoids and questions with NIL answers was at our discretion. In the end we submitted 161 factoids, 30 definitions and nine lists. In previous years we have been obliged to devise a considerable number of temporally restricted questions and this has proved very difficult to do with the majority of them being very

contrived and artificial. For this reason it was intended to set no such questions this year. However, one reasonable one was spotted during the data entry process and so was flagged as such. Two others were also flagged accidentally during data entry. Unfortunately, therefore, the statistics can not tell us anything about temporally restricted questions.

Concerning NIL questions, we have long argued that they tell us very little about the performance of a system unless it can report the reason why there is no answer. For example, this is a useful system:

Q: Who is the Queen of France?

A: France is a Republic!

By contrast, answering NIL would not tell us whether there was an answer which was simply not found, or whether no answer in fact exists. Another important point following from this is that NIL questions artificially boost the performance of a system which returns many NIL answers. For these reasons we decided not to include any questions with NIL answers. However, we would like to see ‘Queen of France’ answers being returned in future workshops.

The grouped nature of the questions had a considerable effect on their difficulty; instead of a series of ‘trivia’ type questions, each with a simple, clear answer, a single topic was effectively investigated in much more detail. To achieve the goals set by the organisers it was necessary to find topics about which several questions could be asked and then to devise as many questions as possible from that topic. Each task was surprisingly hard, and an inevitable consequence was that the questions are much harder this year than in previous years. We had no wish to set especially difficult or convoluted questions, but unfortunately this arose as a side-effect of the new procedures.

The requirement for related questions on a topic necessarily implies that the questions will refer to common concepts and entities within the domain in question. In a series of questions this is accomplished by co-reference – a well known phenomenon within Natural Language Processing which nevertheless has not been a major factor in the success of QA systems at previous CLEF workshops. The most common form is anaphoric reference to the topic declared implicitly in the first question, e.g.:

Q: What is a Polygraph?

Q: When was *it* invented?

However, other forms of co-reference occurred in the questions. Here is an example:

Q: Who wrote the song "Dancing Queen"?

Q: How many people were in *the group*?

Here *the group* refers to the category of entity into which the answer to the first question is known by the questioner to belong. However, the QA system does not know this and has to infer it, a task which can be very complex and indirect, especially where the topic is concealed from the participants.

In addition to the issue of question grouping, it was decided at a very late stage to use not only the two collections from last year (the LA Times and Glasgow Herald) but also the English Wikipedia. The latter is extremely large and greatly increases the task complexity for the participants in terms of both indexing and IR searching. In addition, some questions had to be heavily qualified in order to reduce the ambiguity introduced by alternative readings in the Wikipedia. Here is an example:

Q: What is the "KORG" on which Niky Orellana is a soccer commentator?

Thirdly, we should bear in mind that the Wikipedia varies considerably in size depending on the language, with the English one being by far the largest. We have not controlled for this fact in CLEF and the consequence could be that the addition of Wikipedia had a greater effect on difficulty for English than it did for other languages.

Summary Statistics. Eight cross-lingual runs with English as target were submitted this year, as compared with thirteen for last year. Five groups participated in six source languages, Dutch, French, German, Indonesian, Romanian and Spanish. DFKI submitted runs for two source languages, German and Spanish, while all other groups worked in only one. Cindi Group and Macquarie University both submitted two runs for a language pair

(French-English and Dutch-English respectively) but unfortunately there was no language for which more than one group submitted a run. This means that no direct comparisons can be made between QA systems this year, because the task being solved by each was different.

Assessment Procedure. An XML format was used for the submission of runs this year, by contrast with previous years when fairly similar plain text formats were adopted. This meant that our evaluation tools were no longer usable. However, last year we also participated in the evaluation of the WiQA task organised by University of Amsterdam. For this they developed an excellent web-based tool which was subsequently adapted for this year's Dutch CLEF evaluations. We are extremely grateful to Martin de Rijke and Valentin Jijkoun for allowing us to use it and for setting it up in Amsterdam especially for us. It allows multiple assessors to work independently, shows runs anonymised, allows all answers to a particular question to be judged at the same time (like the TREC software), and includes the supporting snippets for each submitted answer as well as the 'correct' (reference) answer. It also shows inter-assessor disagreement, and, once this has been eliminated, can produce the assessed runs in the correct XML format. Overall, this software worked perfectly for us and saved us a considerable amount of time.

All answers were double-judged. Where assessors differed, the case was discussed between us and a decision taken. We measured the agreement level by two methods. For Agreement 1 we take agreement on each group of 8 answers to a question as a whole as either exactly the same for both assessors or not exactly the same. This is a very strict measure. There were disagreements for 30 questions out of the 200, i.e. 15%, which equates to an agreement level of 85%.

For Agreement Level 2 we taking each decision made on one of the eight answers to a question and count how many decisions were the same for both assessors and how many were not the same. There were 39 differences of decision and a total of 1600 decisions (200 questions by eight runs). This is 2.4%, which equates to an agreement level of 97.6%. This is the measure we used in previous years. Last year the agreement level was 89% and the previous year it was 93%. We conclude from these figures that the assessment of our CLEF runs is quite accurate and that double judging is sufficient.

Results Analysis. As in previous years there were three types of question within the question groups, Factoids, Definitions and Lists. Considering all question types together, the best performance is University of Wolverhampton with 28 R and 2 X, (14% strict or 15% lenient) closely followed by the CINDI Group at Concordia University with 26 R and 1 X (13% strict or 13.50% lenient). Note that these systems are working on different tasks (RO-EN and FR-EN respectively) as noted above, so the results are not directly comparable. The best performance last year for English targets was 25.26%. Nevertheless, considering the extreme difficulty of the questions, this represents a remarkable achievement for these systems.

For Factoids alone, the best system was CINDI (FR-EN) at 11.18% followed by University of Indonesia (IN-EN) with 10.56%. For Definitions the best result was University of Wolverhampton (RO-EN) with 43.33% correct, followed equally by CINDI (FR-EN) and DFKI (DE-EN) both with 23.33%. It is interesting that this year the best Definition score is almost four times the best Factoid score, whereas last year they were nearly equal. One reason for this may be that the definitions either occurred first in a group of questions or on their own in a 'singleton' group. This was not specifically intended but seems to be a consequence of the relationship between Factoids and Definitions, namely that the latter are somehow epistemologically prior to the former¹. In consequence, Definitions may be more simply phrased than Factoids and in particular may avoid co-reference in the vast majority of cases.

Nine lists questions were set but only CINDI was able to answer any of them correctly (11.11% accuracy). (University of Indonesia was ineXact on one list question.) Perhaps the problem here was recognising the list question in the first place – unlike at TREC they are not explicitly flagged. We believe this is not necessarily reasonable since in a real dialogue a questioner would surely make it quite clear whether they expected a list of answers or just one. They would not come up with a list question out of the blue.

6.3 French as Target

Details can be found in the online version of these Working Notes.

¹ Perhaps it is just a consequence of setting too many undergraduate examination papers!

6.4 German as Target

Two research groups submitted runs for evaluation in the track having German as target language: The German Research Center for Artificial Intelligence (DFKI) and the Fern Universität Hagen (FUHA). Both provided system runs for the monolingual scenario and just DFKI submitted runs for the cross-language English-German and Portuguese-German scenario. The assessment was conducted by two native German speakers with fair knowledge of information access systems. Compared to the previous editions of the evaluation forum, this year a decrease in the accuracy of the best performing system and of an aggregated virtual system for both **monolingual** and **cross-language** tasks was registered.

Table 9: Results through the years.

Year	Best Mono	Aggregated Mono	Best Cross	Aggregated Cross
2007	30	45	18.5	18.5
2006	42.33	64.02	32.98	33.86
2005	43.5	58.5	23	28
2004	34.01	43.65	0	0

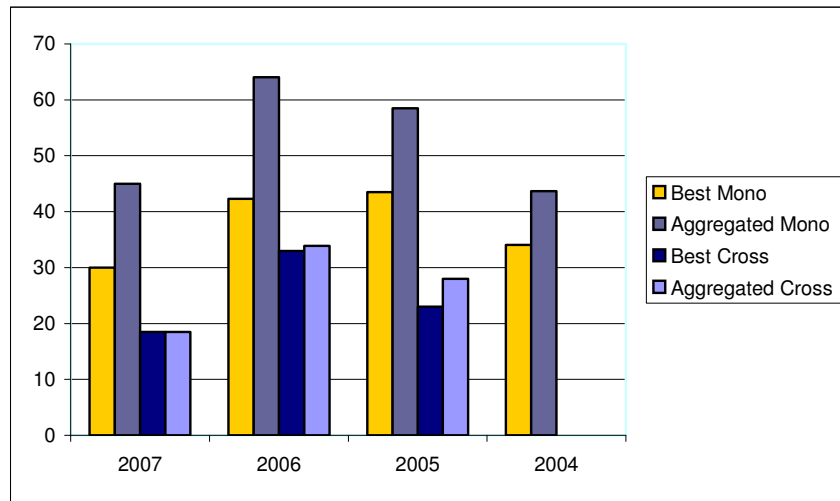


Figure 2: Results evolution

The details of systems' results can be seen in Table 10. There were no NIL questions tested in this year's evaluation. The results submitted by DFKI did not provide a normalized value for the confidence score of an answer and therefore both CWS and K1 values could not be computed.

Table 10. System Performance – Details

Run	R #	W #	X #	U #	% F [164]	% T [27]	% D [28]	% L [8]	NIL		CWS	K1	Overall accuracy
									#	% [0]			
<i>dfki071dede_M</i>	60	121	14	5	29.8	14.81	39.29	0	0	0	0	0	30
<i>fuha071dede_M</i>	48	146	4	2	24.39	18.52	28.57	0	0	0	0.086	-0.17	24
<i>fuha072dede_M</i>	30	164	4	2	17.07	14.81	7.14	0	0	0	0.048	-0.31	15
<i>dfki071ende_C</i>	37	144	18	1	17.68	14.81	25	12,5	0	0	0	0	18.5
<i>dfki071ptdec</i>	10	180	10	0	3.66	7.41	14.29	0	0	0	0	0	5

The number of topics covered by the questions was of 116 distributed as it follows: 69 topics consisting of 1 question, 19 topics of 2 related questions, and each 19 topics of 3 and 4 related questions. The most frequent topic types were PERSON (40), OBJECT (33) and ORGANIZATION (23). As regards the source of the

answers, 101 questions from 68 topics asked for information out of the CLEF document collection and the rest of 99 from 48 topics for information from Wikipedia. The distribution of the topics over the document collections (CLEF vs. Wikipedia) is as follows: 53 vs. 16 topics of 1 question, 4 vs. 15 topics of each 2 and 3 questions and 7 vs. 2 topics of 4 questions.

Table 11: Inter-Assessor Agreement/Disagreement (breakdown)

Run ID	# Questions	# Q-Disagreements						
		Total	F	D	L	X	U	W/R
<i>dfki071dede_M</i>	200	20	16	4	0	15	4	1
<i>fuha071dede_M</i>	200	13	10	3	0	7	3	3
<i>fuha072dede_M</i>	200	7	6	1	0	2	2	3
<i>dfki071ende_C</i>	200	13	7	5	1	12	1	0
<i>dfki071ptde_C</i>	200	8	3	5	0	8	0	0

Table 12 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. Along the total figures for the 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 answer 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.

6.5 Italian as Target

Only one group took part in this year to the monolingual Italian task, i.e. FBK-irst, achieving the following results:

Table 12: Results.

Run	R	W	X	U	% F	% T	% D	% L]	NIL		CWS	K1	Overall accuracy
									Returned	Correct			
<i>irst071ITIT</i>	23	160	4	13	15.17	12.5	2.63	0	14	3	0.0165	-0.0429	11,55

More details on Italian as a target can be found in the online version of this working notes.

6.6 Portuguese as Target

Six research groups took part in tasks with Portuguese as target language, submitting eight runs: seven in the monolingual task, and one with English as source; unlike last year, no group presented Spanish as source. One new group (INESC) participated this. The group of University of Évora (UE) returned this year, while the group from NILC, the sole Brazilian group to take part to date, was absent.

Again, Priberam presented the best result for the third year in a row; the group of the University of Évora wasn't however far behind. As last year, we added the classification X-, meaning incomplete, while keeping the classification X+ for answers with extra text or other kinds of inexactness. In Table 3 we present the overall results.

Table 13: Results of the runs with Portuguese as target: all 200 questions

Run Name	R (#)	W (#)	X+ (#)	X- (#)	U (#)	Overall Accuracy (%)	NIL Accuracy	
							Precision (%)	Recall (%)
diue071ptpt	84	103	1	11	1	42.0	11.7	92.3
esfi071ptpt	16	178	0	4	2	8.0	6.3	69.2
esfi072ptpt	12	184	0	2	2	6.0	6.1	84.6
feup071ptpt	40	158	1	1	0	20.0	8.3	84.6
ines071ptpt	22	171	1	4	2	11.0	7.3	69.2
ines072ptpt	26	168	0	4	2	13.0	7.2	84.6
prib071ptpt	101	88	5	5	1	50.5	27.8	46.2
lcc_071enpt	56	121	7	3	13	28.0	33.3	23.1

A direct comparison with last year's results is not fully possible, due to the existence of multiple questions to each topic. Therefore, 14 presents results regarding the first question of each topic, which we believe is more readily comparable to the results of previous years.

Table 14: Results of the runs with Portuguese as target: answers to the first question of the 149 topics

Run Name	R (#)	W (#)	X+ (#)	X- (#)	U (#)	Overall Accuracy (%)
diue071ptpt	61	77	1	9	1	40,9%
esfi071ptpt	11	132	0	4	2	7,4%
esfi072ptpt	6	141	0	1	1	4,0%
feup071ptpt	34	113	1	1	0	22,8%
ines071ptpt	17	125	1	4	2	11,4%
ines072ptpt	21	122	0	4	2	14,1%
prib071ptpt	92	86	3	5	1	61,7%
lcc_071enpt	44	48	7	3	9	29,5%

As it can be seen, the removal of subsequent questions to each topic doesn't cause a big change on the overall results, apart from a clear improvement by Priberam. On the whole, compared to last year (Vallin et al., 2007), Priberam saw a slight drop on its results, Raposa (FEUP) a clear improvement from an admittedly low level, Esfinge (SINTEF) a clear drop, and LCC kept last year's levels. Senso (UE) shows a marked improvement since its last participation in 2005. We leave it to the participants to comment on whether it might have been caused by harder questions or changes (or lack thereof) in the systems.

Question 94 was reclassified as NIL due to a spelling error, and question 135 because of the use of a word with a rare meaning. On the other hand, one system saw through that rare meaning, providing a correct answer; we decided to keep the question as NIL, considering correct both the system's answer and any NIL answer from other systems. The same system also found a correct answer to a NIL question, not discovered during the question creating process; that question was therefore reclassified as non-NIL. In the end, there were 13 NIL questions.

Table 15 shows the results for each answer type of definition questions, while Table 16 shows the results for each answer type of factoid questions (including list questions). As it can be seen, four out of six systems perform clearly better when it comes to definitions than to factoids. This may well have been helped by the use of Wikipedia texts, where a large proportion of articles begin with a definition.

Table 15: Results of the assessment of the monolingual Portuguese runs: definitions

Run	obj	org	oth	per	TOT	%
	6	6	9	9	30	
diue071ptpt	6	4	5	4	19	63%
esfi071ptpt	1	0	0	0	1	3%
esfi072ptpt	1	0	0	0	1	3%
feup071ptpt	3	2	4	7	16	53%
ines071ptpt	4	4	6	0	14	47%
ines072ptpt	5	5	6	2	18	60%
prib071ptpt	6	4	6	7	23	77%
<i>combination</i>	6	5	8	9	27	87%
lcc_071enpt	2	3	2	1	8	27%

Table 16: Results of the assessment of the Portuguese runs: factoids, including lists

Run	cou	loc	mea	obj	org	oth	per	tim	TOT	%
	21	31	16	5	21	26	21	19	160	
diue071ptpt	11	17	4	3	6	8	7	9	65	39%
esfi071ptpt	3	3	0	0	1	0	1	7	15	9%
esfi072ptpt	2	4	0	0	1	0	2	2	11	7%
feup071ptpt	4	8	0	0	3	1	3	5	24	15%
ines071ptpt	1	3	0	0	0	0	2	2	8	5%
ines072ptpt	2	4	0	0	0	0	2	2	10	6%
prib071ptpt	9	15	10	1	11	14	8	10	78	46%
<i>combination</i>	16	24	12	3	12	17	12	13	109	68%
lcc_071enpt	7	11	6	1	3	10	4	6	48	29%

We included in both Table 15 and in Table 16 a virtual run, called combination, in which one question is considered correct if at least one participating system found a valid answer. The objective of this combination run is to show the potential achievement when combining the capacities of all the participants. The combination run can be considered, somehow, state-of-the-art in monolingual Portuguese question answering. The system with best results, Priberam, answered correctly 72.7% the questions with at least one correct answer, not as dominating as last year. Despite being a bilingual run, LCC answered correctly 14 questions not answered by any of the monolingual systems.

In Table 17, we present some values concerning answer and snippet size (in number of words).

Table 17: average size of answers

Run name	Non-NIL Answers (#)	Average answer size	Average answer size (R only)	Average snippet size	Average snippet size (R only)
diue071ptpt	89	2.8	2.9	25.0	24.3
esfi071ptpt	57	2.4	2.8	56.3	29.3
esfi072ptpt	19	2.4	2.8	59.7	29.1
feup071ptpt	56	2.7	3.3	59.8	32.9
ines071ptpt	49	3.7	4.8	60.7	33.6
ines072ptpt	47	3.8	5.3	61.7	34.2
prib071ptpt	182	3.5	4.4	49.6	32.4
lcc_071enpt	191	3.4	4.2	45.2	32.7

Temporally restricted questions: Table 18 presents the results of the 20 temporally restricted questions. As in previous years, the effectiveness of the systems to answer those questions is visibly lower than for non-TRQ questions (and indeed several systems only answered correctly question 160, which is a NIL TRQ).

Table 18: accuracy of temporally restricted questions

Run name	Correct answers (#)	T.R.Q correctness (%)	Non-T.R.Q correctness (%)	Total correctness (%)
diue071ptpt	4	20.0	44.4	42.0
esfi071ptpt	1	5.0	8.3	8.0
esfi072ptpt	1	5.0	6.1	6.0
feup071ptpt	1	5.0	21.7	20.0
ines071ptpt	1	5.0	11.7	11.0
ines072ptpt	1	5.0	15.0	14.0
prib071ptpt	8	40.0	51.7	28.0
lcc_071enpt	6	30.0	27.8	50.5

List questions: a total of twelve questions were defined as list questions; unlike last year, all these questions were closed list factoids, with two to twelve answers each². The results were, in general, weak, with UE and LCC getting two correct answers, Priberam five, and all other system zero. There was a single case of incomplete answer (i.e., answering some elements of the list only), but it was judged W since, besides incomplete, it was also unsupported.

6.7 Romanian as Target

At CLEF 2007 Romanian was addressed as a target language for the first time, based on the collection of Wikipedia Romanian pages frozen in November 2006, and as a source language for the second time, using the English news collection (Los Angeles Times, 1994 and Glasgow Herald, 1995) and the Wikipedia English pages.

Creation of Questions. The creation of the questions was realized at the Faculty of Computer Science, A.I. Cuza University of Iasi. The group³ was very well instructed with respect to this task, using the Guidelines for Question Generation and based on a good feedback received from the organizers at IRST⁴. The final 200 created questions are distributed according to table 19.

Table 19: Question types distribution in Romanian

	PERSON	TIME	LOCATION	ORGA-NIZATION	MEASURE	COUNT	OBJECT	OTHER	TOTAL
FACTOID	22	17	21	19	17	20	16	21	153
DEFINITION	9			5			6	10	30
LIST	10								10
NIL QUESTIONS	7								7

Most difficulties in this task were raised by deciding on the supporting snippets, especially for questions belonging to the same topic. We found unnatural to include answers through “copy-paste” from the text: if the question requires an answer in the Nominative case, but the text includes the answer in the Genitive case, then we had to include the Genitive in the answer, even if it is more natural to have the answer in Nominative.

Participants. This year two Romanian groups took part in the monolingual task with Romanian as a target language: the Faculty of Computer Science from the A.I. Cuza University of Iasi, and the Research Institute for

² There were some open list questions as well, but they were classified and evaluated as ordinary factoids.

³ Three Computational Linguistics Master students: Anca Onofraş, Ana-Maria Rusu, Cristina Despa, supervised and working in collaboration with the two organizers

⁴ Without the help received from Danilo Giampiccolo and Pamela Forner, we wouldn’t have solved all our problems.

Artificial Intelligence from the Romanian Academy, Bucharest. Three runs were submitted – one by the first group and two by the second group, with the differences between them due to the way they treated the question-processing and the answer-extraction. The 2007 results are presented in Tables 20 below. One system with Romanian as a source language and English as target was submitted by the Computational Linguistics Group from the University of Wolverhampton, United Kingdom.

Tables 20: Results in the monolingual task, Romanian as target language

Run	R	W	X	U	Overall Accuracy	NIL RETURNED	NIL correct	CWS
outputRoRo (1)	24	171	4	1	12	100	5	0.02889
ICIA071RORO (2)	60	105	34	1	30	54	7	0.09522
ICIA072RORO (3)	60	101	39	0	30	54	7	0.09522

All three systems “crashed” on the LIST questions. The NIL questions are hard to classify, starting from the question-classifier (the classifier should “know” that the QA system has no possibility, no knowledge to find the answer). It would be better to have a clear separation between the NIL answers due to impossibility to find answer and the NIL answers classified as such by the system. None of the three systems could handle the questions related under one same topic: the systems returned at most the answer to the first question in a topic.

Assessment Procedure. Due to time restrictions, all three runs were judged by only one assessor at the Faculty of Computer Science in Iasi, so an inter-annotator agreement was not possible. Based on the Guidelines, all three systems were judged in parallel. The same evaluation criteria, especially with respect to the UNSUPPORTED and INEXACT answers, were used.

6.8 Spanish as Target

The participation at the Spanish as Target subtask has decreased from 9 groups in 2006 to 5 groups this year. All the runs were monolingual. We think that the changes in the task (linked questions and wikipedia) led to a lower participation and worse overall results because systems could not be tuned on time. Table 21 shows the summary of systems results with the number of Right (R), Wrong (W), Inexact (X) and Unsupported (U) answers. The table shows 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. All the runs were assessed by two assessors. Only a 1.5% of the judgements were different and the resulting kappa value was 0,966, which corresponding to “almost perfect” assessment [7].

Table 21: Results at the Spanish as target.

Run	R #	W #	X #	U #	% F [115]	% T [43]	% D [32]	% L [10]	NIL		CWS	KI	Overall accuracy %
									#	F [8]			
Priberam	89	87	3	21	47,82	23,25	68,75	20	3	0,29	-	-	44,5
Inaoe	69	118	7	6	28,69	18,60	87,50	-	3	0,12	0,175	-0,287	34,5
Miracle	30	158	4	8	20	13,95	3,12	-	1	0,07	0,022	-0,452	15
UPV	23	166	5	6	13,08	9,30	12,5	-	1	0,03	0,015	-0,224	11,5
TALP	14	183	1	2	6,08	2,32	18,65	-	3	0,07	0,007	-0,34	7

Best performing systems have obtained worse results than last year due mainly to the low performance in answering linked questions (15% of the questions) and due to the questions with answer only in Wikipedia. Table 22 shows that considering only self-contained questions (the first one of each topic group) the results are closer to the ones obtained last year. In fact the accuracy for the linked questions is less than 20%.

Table 22. Results for self-contained and linked questions, compared with overall accuracy.

Run	% accuracy over Self-contained questions [170]	% accuracy over Linked questions [30]	% Overall Accuracy [200]
Priberam	49,41	16,66	44,5
Inaoe	37,64	16,66	34,5
Miracle	15,29	13,33	15
UPV	12,94	3,33	11,5
TALP	7,05	6,66	7

Table 23 shows some evidence on the effect of Wikipedia in the performance. When the answer appears only in Wikipedia the accuracy is reduced in more than 35% in all the cases.

Table 23: Results for questions with answer in Wikipedia

Run	% accuracy over questions with answer only in wikipedia [114]	% accuracy over questions with answer in both EFE and wikipedia [71]
Priberam	40.35%	54.93%
Inaoe	29.82%	42.25%
Miracle	7.89%	28.17%
UPV	7.02%	19.72%
TALP	0%	14.08%

Regarding NIL questions, Table 24 shows the harmonic mean (F) of precision and recall for self-contained, linked and all questions. The best performing system has decreased their overall performance with respect to the last edition (see Table 25) in NIL questions. However, the performance considering only self-contained questions is closer to the one obtained last year.

Table 24: Results at the Spanish as target for NIL questions

	F-measure (Self- contained)	F- measure (Overall)	Precision (Overall)	Recall (Overall)
Priberam	0.4	0.29	0.23	0.38
Inaoe	0.13	0.12	0.07	0.38
Miracle	0.07	0.07	0.05	0.13
UPV	0.04	0.03	0.02	0.13
TALP	0.06	0.07	0.04	0.38

Table 25. Evolution of best results in NIL questions.

Year	F-measure
2003	0,25
2004	0,30
2005	0,38
2006	0,46
2007	0,29

The correlation coefficient r between the self-score and the correctness of the answers (shown in Table 26) has been similar to the obtained last year, being not good enough yet, and explaining the low results in CWS and K1 [6] measures.

Since a supporting snippet is requested in order to assess the correctness of the answer, we have evaluated the systems capability to extract the answer when the snippet contains it. The first column of table 27 shows the percentage of cases where the correct answer was present in the snippet and 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 process. As shown in the table, the best systems are also better in the task of answer extraction, whereas the rest of systems still have a lot of room for improvement.

Table 26. Answer Extraction and correlation coefficient r results at the Spanish as target

Run	% Answer Extraction	r
Priberam	93,68	-
INAOE	75	0,1170
Miracle	49,18	0,237
UPV	54,76	-0,1003
TALP	53,84	0,134

7 Final considerations

This year the task was changed considerably and this affected the general level of results and also the level of participation in the task. The grouped questions could be regarded as more realistic and more searching but in consequence they were much more difficult. The policy of not declaring the question type means that if this is deduced incorrectly then the answer is bound to be wrong. Moreover, the policy of not even declaring the topic of a question group, but leaving it implicit (usually within the first question) means that if a system infers the topic wrongly, then all questions in the group will be answered wrongly. This should be probably re-considered, as it is not 'realistic'. In a real dialogue, if a question is answered inappropriately we do not dismiss all subsequent answers from that person, we simply re-phrase the question instead. The level of ambiguity concerning question type in a real dialogue is not fixed at some arbitrary value but varies according to many factors which the questioner estimates. In CLEF we are not modelling this process at all accurately and this affects the validity of our results. Finally, co-reference has now entered CLEF. This is interesting and useful but it might be preferable if we could separate the effect of co-reference resolution from other factors in analysing results. This could be done by marking up the co-references in the question corpus and allowing participants to use this information under certain circumstances.

Acknowledgments

A special thank to Bernardo Magnini (FBK-irst, Trento, Italy), who has given his precious advise and valuable support at many levels for the preparation and realization of the QA track at CLEF 2007.

Anselmo Peñas has been partially supported by the Spanish Ministry of Science and Technology within the Text-Mess-INES project (TIN2006-15265-C06-02).

References

1. QA@CLEF website: <http://clef-qa.itc.it/>
2. AVE Website: <http://nlp.uned.es/QA/ave/>.
3. QAST Website: <http://www.lsi.upc.edu/~qast/>
4. QART Website: <http://gplsi.dlsi.ua.es/qart/>
5. QA@CLEF 2007 Organizing Committee. Guidelines 2007. http://clef-qa.itc.it/2007/download/QA@CLEF07_Guidelines-for-Participants.pdf
6. Herrera, J., Peñas A., Verdejo, F.: Question answering pilot task at CLEF 2004. In: Peters, C., Clough, P., Gonzalo, J., Jones, Gareth J.F., Kluck, M., Magnini, B. (eds.): Multilingual Information Access for Text, Speech and Images. Lecture Notes in Computer Science, Vol. 3491. Springer-Verlag, Berlin Heidelberg New York (2005) 581–590
7. J. R. Landis and G. G. Koch. 1997. The measurements of observer agreement for categorical data. *Biometrics*, 33:159–174.