Overview of the CLEF 2004 Multilingual Question Answering Track

Magnini, B.; Vallin, A.; Ayache, C.; Erbach, G.; Penas, A.; de Rijke, M.; Rocha, P.; Simov, K.; Sutcliffe, R.

Published in:
Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign

Overview of the CLEF 2004 Multilingual Question Answering Track

Bernardo Magnini[1], Alessandro Vallin[2], Christelle Ayache[3], Gregor Erbach[4], Anselmo Peñas[5], Maarten de Rijke[6], Paulo Rocha[7], Kiril Simov[8] and Richard Sutcliffe[9]

Abstract. Following the pilot Question Answering Track at CLEF 2003, a new evaluation exercise for multilingual QA systems took place in 2004. This paper reports on the novelties introduced in the new campaign and on participants' results. Almost all the cross-language combinations between nine source languages and seven target languages were exploited to set up more than fifty different tasks, both monolingual and bilingual. New types of questions (How-questions and definition questions) were given as input to the participating systems, while just one exact answer per question was allowed as output. The evaluation exercise has highlighted some difficulties in assessing definition questions and can be improved in the future, but the overall analysis of submissions shows encouraging results.

1 Introduction

Question Answering (QA) systems have been evaluated for the last six years at the TREC campaigns. The TREC QA tracks have evolved over the years, so that increasingly difficult tasks have been proposed, addressing not only factoid but also list and definition questions, and requiring exact answers instead of longer text snippets as output [8]. Nevertheless, multilinguality has never been investigated at TREC's QA track, thus leaving room for challenging tasks in languages other than English or even across different languages, which is actually in the focus of the CLEF campaigns.

The first multilingual QA track at CLEF took place in 2003. Eight groups from Europe, the U.S. and Canada participated in nine tasks, submitting a total of seventeen

[1] ITC-Irst, Trento, Italy (magnini@itc.it).
[2] ITC-Irst, Trento, Italy (vallin@itc.it).
[3] ELDA/ELRA, Paris, France (ayache@elda.fr).
[4] DFKI, Saarbrücken, Germany (erbach@dfki.de).
[5] Departamento de Lenguajes y Sistemas Informáticos, UNED, Madrid, Spain (anselmo@lsi.uned.es).
[6] Informatics Institute, University of Amsterdam, The Netherlands (mdr@science.uva.nl).
[7] Linguateca, Braga Node, Universidade do Minho, Potugal (Paulo.Rocha@alfa.di.uminho.pt).
[8] IPP, Bulgarian Academy of Sciences, Sofia, Bulgaria (kivs@bultreebank.org) .
[9] DLTG, University of Limerick, Ireland (richard.sutcliffe@ul.ie).

runs. Three languages were addressed in the monolingual tasks (Dutch, Italian and Spanish), while in the bilingual tasks questions were formulated in five source languages (Dutch, French, German, Italian and Spanish) and searched for answers in an English document collection. It was a pilot evaluation exercise and 200 simple, fact-based questions were given as input in all tasks, and participants were allowed to return up to three responses per questions, either exact or 50 bytes-long answer-strings [6].

In 2004 the QA@CLEF track[10] attracted considerable attention within the CLEF framework; in fact three different tasks were devoted to it: the main QA track, a Spanish pilot task and iCLEF, the interactive track. The main track has included more European languages and all the cross-language combinations between them have been exploited in order to set up a number of different tasks. As a result, the CLEF QA community has grown and eighteen groups tested their systems, submitting forty-eight runs.

This paper provides an overview of the main QA track. The following sections report on the languages considered in the experiments, on the procedure that was adopted to build the test sets, and on the participants' results. Each target language will be treated separately, as different subtasks.

2 Tasks

Though Chinese has the highest number of speakers in the world, English has become a sort of lingua franca, as the fact that most web pages world wide are in English testifies. Nevertheless, a lot of information is available in other European languages, among which Spanish, German, French, Italian and Portuguese are the most prominent. This motivates the study of multilingual information access.

In a multilingual QA task two main variables need to be considered: the source language, i.e. the language in which the questions are formulated, and the target language, i.e. the language of the document collection. A cross-language QA system should enable users to search documents that are written in a language they do not know, which is a promising application in a multilingual society. Answer-strings, which are usually retrieved from the corpus without any changes, could be translated into the source language, but this further cross-lingual step was not required in the track.

2.1 Languages

In 2004 nine source languages (Bulgarian, Dutch, English, Finnish, French, German, Italian, Portuguese and Spanish) and seven target languages (Dutch, English, French, German, Italian, Portuguese and Spanish) were considered at the CLEF QA track. Almost all combinations between source and target have been exploited in order to propose as many tasks as possible: since no document collections were available, Bulgarian and Finnish were considered as source languages only, while the

[10] URL: http://clef-qa.itc.it/2004/

monolingual English task was discarded because it has been "traditionally" in the focus of the TREC campaigns. A total of 56 tasks were set up, divided into 6 monolingual (Dutch, French, German, Italian, Portuguese and Spanish) and 50 bilingual.

2.2 The Evaluation Exercise

Since QA systems process natural language questions rather than keywords and retrieve precise answers rather than entire documents, 200 questions were provided as input in all the tasks, and exact answer-strings were required as output.

The target corpora in all the languages were collections of newspapers and news agencies' articles. The texts were SGML tagged, and each document had a unique identifier (docid) that systems had to return with the answer, in order to support it. The corpora, released by ELRA/ELDA, were large, unstructured, open-domain text collections.

The 200 questions given as input in the tasks were fact-based, but about the 10% of each test set was made up of definition questions such as *What is UNICEF?* or *Who is Tony Blair?*, which were not included in 2003. In addition, another 10% did not have any answer in the corpora, and the right response to those questions was the string "NIL".

Each target language had its own 200 questions and, despite the efforts of the co-ordinators, there was just a little overlap between the test sets of different target languages: just two questions recurred in all the test sets and, on average, each test set shared 10 questions with the other ones.

As far as the answers are concerned, the requirements were stricter than in 2003, when participants were allowed to submit either exact or 50 bytes-long answers. Due to the potential number of participants attracted by so many tasks, the evaluation efforts needed to be minimised, so in 2004 the output was reduced to a single, exact answer-string.

Generally speaking, on the one hand the QA track at CLEF 2004 tried to attract as many participants as possible in a non-competitive setting, while on the other hand the co-ordinators aimed at reflecting the development that the TREC tracks have been undergoing over the years. For this reason, the guidelines reflected to a large extent those of the TREC 2002 QA track, adopting similar requirements and evaluation measures.

3 Test Set Preparation

Multilingual QA entails a number of subtasks, such as the development of tools (PoS-taggers, parsers and Named Entity recognisers) for languages other than English and the translation of questions and answers into other languages [1]. The construction of a reusable, multilingual collection of questions with the related [answer-strings, docid] pairs represents a useful resource, and the CLEF QA evaluation exercise offers the opportunity to create such a benchmark. As in the 2003 campaign, when two multilingual Gold Standard collections of questions and answers were built [5 and 6],

in 2004 the generation of the test sets was closely monitored and exploited in order to build similar test sets for all the tasks, and to translate all the questions proposed into the track in all the source languages. Because of the number of languages involved, there was no attempt to have exactly the same test set in all the tasks, as we managed to do in 2003.

Eight groups were involved in the generation, translation and manual verification of the questions: the IPP group at the Bulgarian Academy of Sciences translated the entire collection of questions and answers in Bulgarian, DFKI created the German test set, ELRA/ELDA took over the work on the French questions, ITC-Irst was in charge of the Italian and English test sets, Linguateca provided the Portuguese part of the benchmark, UNED prepared the Spanish part, the University of Amsterdam worked on Dutch and the University of Helsinki joined the activity translating 200 English questions into Finnish, in order to set up the Finnish-English task.

3.1 Question Generation

The questions in the test sets addressed large (on average 230 Mb), open domain corpora. The document collections for all the target languages were comparable because they were made up of newspapers and news agencies articles that referred to the same time-span: *NRC Handelsblad* (years 1994 and 1995) and *Algemeen Dagblad* (1994 and 1995) for Dutch; *Los Angeles Times* (1994) and *Glasgow Herald* (1995) for English; *Le Monde* (1994) and *SDA French* (1994 and 1995) for French; *Frankfurter Rundschau* (1994), *Der Spiegel* (1994 and 1995) and *SDA German* (1994 and 1995) for German; *La Stampa* (1994) and *SDA Italian* (1994 and 1995) for Italian; *PÚBLICO* (1994 and 1995) for Portuguese and *EFE* (1994 and 1995) for Spanish.

As a first step in the test sets preparation, each co-ordinating group generated 100 questions in its own target language, searched manually for at least one answer per question supported by a document and then translated into English, that was used as the interlingua between all the groups, both questions and answers. The questions had to be compliant with specific criteria that were previously established: list questions (e.g. *What are the three most important export products of Italy?*), embedded questions (e.g. *When did the king who succeeded Queen Victoria die?*), yes/no questions (e.g. *Did Shakespeare have any sisters?*) and Why- questions (e.g. *Why did Nixon resign?*) were not considered in the track [2].

On the other hand, the test set included two question types that were avoided in 2003: How- questions and definition questions. These two categories, which can have longer answer-strings than the factoid questions, were approached basically in the same way, though assessors were less demanding in terms of exactness.

How- questions (e.g. *How did Hitler die?*), may have several different responses (e.g. *He committed suicide*, or *in mysterious circumstances* or *hit by a bullet*, or even *alone*) that provide different kinds of information.

Similarly, definition questions (e.g. *What is the atom?* or *Who are the Martians?*) are considered very difficult because though their target is clear, they are posed in isolation, and different questioners might expect different answers depending on their previous assumptions. They were first introduced at TREC 2001 and then

proposed again in 2003, when organisers tried to define a potential user of the QA system, who would be "an adult, a native speaker of English, and an 'average' reader of US newspapers" [8]. TREC assessors created a list of "information nuggets" (i.e. significant facts that were likely to appear in the desired response), some of which were necessary, and judged the content of each answer checking how many nuggets it contained. This way of assessing the definition questions was quite complex and far from being exhaustive, so the CLEF approach in this sense has been simplified: first of all only definition questions that referred to either a person or an organisation were chosen, in order to avoid more abstract "concept definition" questions such as *What is religion?*, that would be too complex to be judged. The restriction to persons (*Who is Kofi Annan?*) and organisations (*What is Amnesty International?*) aimed at generating simple definition questions, whose answer could be a single, well defined text snippet such as *British spies listened in to UN Secretary General Kofi Annan's office* or *Amnesty International campaigns for human rights*, without any previous expectations regarding the most relevant information that a system should return. Secondly, as they were introduced as a stepping stone in 2004, the most general answers were judged as correct, assuming that potential users did not know anything about the addressed person or organisation.

The track co-ordinators attempted to balance the test sets according to the different answer types of the questions. Eight answer types were considered: TIME (e.g. *What year was Thomas Mann awarded the Nobel Prize?*), MEASURE (e.g. *How many years of imprisonment did Nelson Mandela serve?*), PERSON (e.g. *Who was Lisa Marie Presley's father?*), ORGANISATION (e.g. *What is the name of the Kurdish separatist party?*), LOCATION (*What is the capital of Japan?*), OBJECT (e.g. *Name an odourless and tasteless liquid.*), MANNER (e.g. *How did Pasolini die?*) and OTHER (e.g. *What animal coos?*). It is difficult to determine the intrinsic difficulty of a question, but the distribution of several answer types in the test sets could differentiate the task and offer some insights in the systems performance with regard to particular categories of questions, as we will show in the results section below.

Each organising group (except IPP and the University of Helsinki) collected 100 questions that had at least one answer in their own target corpus. Those questions would be shared with the other groups, so they were translated into English and saved in a simple XML format. For instance, during this work phase ELRA/ELDA generated the factoid question *Où se trouve Halifax ?*, that had a LOCATION as answer type, translating it into *Where is Halifax located?*.

3.2 Translation

Seven hundred questions were formulated in an original source language, manually verified against a document collection, translated into English and collected in a common XML format. In order to share them in a multilingual scenario, a second translation in all the nine source languages of the track was necessary. Native speakers of each source language with a good command of English were recruited, and they were asked to translate the questions trying to adhere as much as possible to the English version. In case of any discrepancies between the original and the English

form, they were expected to follow the former, and to communicate the changes that the latter presented. Nevertheless, cultural differences made some cross-lingual obstacles unavoidable: so, for example, the English question *What does a luthier make?* became tautological in German (*Was macht ein Geigen- und Gitarrenbauer?*), while some other concepts, such as *CEO*, were ambiguous and were translated in different ways (*chairman*, *managing director* or *president*). Moreover, translators encountered difficulties in the transliteration of proper names: for instance, *Vladimir Zhirinovsky* is written *Wladimir Schirinowski* in German, *Vladimir Zhirinovskij* in Italian and *Vladimir Jirinovski* in French. Translators usually chose the most frequent form in which proper names appeared in their target corpus.

Finally, in carrying out the assessments it became clear that translation has a discernible effect on the integrity of the judgement process. For example is a *Finance Minister* the same as a *Minister for Economic Affairs*? These might be (and in fact are) different roles but they could equally be the same one translated differently. Similarly, when is a *General Manager* the same as a *Secretary General*? In English a General Manager is quite a junior managerial position so the answer is probably "never". However in another language they might be quite equivalent. It is hard therefore to know what to conclude from judgements relating to questions describing translated versions of ranks, titles and so on.

In order to reduce inconsistencies, questions were translated into the form in which a native speaker would *naturally* ask it. The fact that manual translation captured some of the cross-cultural as well as cross-language problems is good since QA systems are designed to work in the real world.

3.3 Gold Standard

Once all the 700 questions were translated into eight languages (Finnish was added only shortly before the beginning of the experiments, and just for 200 questions), 100 additional questions for each target language were selected from the collection, in order to collect 200 questions per test set.

Around twenty of them did not have any answer in the document collections, and the right response to them was the string "NIL". The organisers decided not to include any NIL question among the definitions. The usual procedure to choose them was to select those containing proper nouns that did not occur in the document collection. Though it was easy to implement, this strategy probably made it too easy for participating systems to identify NIL questions, and should be reconsidered for future campaigns. Being aware of this drawback, some groups randomly selected the required NIL questions from those that seemed to have no answer in the document collections, and double checked them.

Additional questions were manually verified and new answers were added to those that were just the translation of the original one. Figure 1 below shows a sample from the multilingual collection of questions and answers built by the organising groups, called *Multieight-04 corpus*. From this XML file the plain text test sets used for the evaluation exercise were extracted. Each question is described according to its category (either factoid or definition) and to its answer type. The information

concerning the category was kept also in test sets released to participants, where the character *F* designated a factoid, and *D* a definition. Questions appear in eight languages, and in one or more of them at least one [answer-string, docid] pair is given. The Boolean attribute "original" keeps track of the language in which each question was first generated and verified.

```
<q cnt="0504" category="F" answer_type="LOCATION">
  <language val="BG" original="FALSE">
    <question group="BTB">Къде се намира Халифакс?</question>
    <answer n="1" docid="">TRANSLATION[Канада]</answer>
  </language>
  <language val="DE" original="FALSE">
    <question group="DFKI">Wo liegt Halifax?</question>
    <answer n="1" docid="">TRANSLATION[Kanada]</answer>
  </language>
  <language val="EN" original="FALSE">
    <question group="ELDA">Where is Halifax located?</question>
    <answer n="1" docid="">TRANSLATION[Canada]</answer>
    <answer n="2" docid="LA112094-0062">Canada</answer>
  </language>
  <language val="ES" original="FALSE">
    <question group="UNED">¿Dónde se encuentra Halifax?</question>
    <answer n="1" docid="">TRANSLATION[Canadá]</answer>
    <answer n="2" docid="EFE19940927-15402">Canadá</answer>
  </language>
  <language val="FR" original="TRUE">
    <question group="ELDA">Où se trouve Halifax ?</question>
    <answer n="1" docid="ATS.950616.0005">Canada</answer>
  </language>
  <language val="IT" original="FALSE">
    <question group="IRST">Dove si trova Halifax?</question>
    <answer n="1" docid="">TRANSLATION[Canada]</answer>
  </language>
  <language val="NL" original="FALSE">
    <question group="UoA">Waar is Halifax?</question>
    <answer n="1" docid="">TRANSLATION[Canada]</answer>
  </language>
  <language val="PT" original="FALSE">
    <question group="LING">Onde fica Halifax?</question>
    <answer n="1" docid="">TRANSLATION[Canadá]</answer>
    <answer n="2" docid="LING-940526-150">West Yorkshire</answer>
    <answer n="3" docid="LING-941009-021">Nova Escócia, no Canadá</answer>
    <answer n="4" docid="LING-941201-050">Canadá</answer>
  </language>
</q>
```

Fig. 1. Sample of the *Multieight-04* collection of questions and answers

The entire collection is made up of 608 factoid and 92 definition questions, and the eight answer types are rather balanced: it includes 173 PERSON, 118 LOCATION, 98 ORGANISATION, 88 OTHER, 84 MEASURE, 82 TIME, 31 OBJECT and 26 MANNER. Each question has at least one answer in one or more

target document collections, but due to the variety of languages, just a few were manually verified in all the languages and consequently appeared in all the test sets.

Similar to the *DISEQuA* and the *Multisix* collections built for the CLEF 2003 QA track, *Multieight-04* is a valuable and reusable benchmark resource that can be further enlarged and distributed. Unfortunately it does not contain all the responses to each question, but just those that were manually found for the test sets preparation. It could be enriched with automatically retrieved pattern sets of correct answers in all the languages.

4 Participants

The encouraging results of the 2003 campaign, which led to the consolidation of the CLEF QA community, and probably the variety of the proposed tasks, gave rise to an increase in the number of participating teams. At the CLEF 2003 QA track 8 groups (3 from the U.S. and 5 from Europe) submitted a total of 17 runs in 9 tasks, while in 2004 18 teams (all of them from Europe except one from Mexico) returned 48 runs distributed over 19 monolingual and bilingual tasks. These figures are similar to those of the TREC-8 pilot QA evaluation exercise, where 20 groups submitted 46 runs, and represent a promising starting point for future campaigns, in which participants from other parts of the world should be involved.

Table 1. The tasks and the corresponding number of submitted runs at the CLEF 2004 QA track

Source Languages \ Target Languages	DE	EN	ES	FR	IT	NL	PT
BG		1		2			
DE	2	3		2			
EN				2		1	
ES			8	2			
FI		1					
FR		6		2			
IT		2		2	3		
NL				2		2	
PT				2			3

As Table 1 shows, many of the 56 tasks that were set up did not attract any participants, but in all the six monolingual tasks, highlighted in the table with grey cells, two or more runs were returned. Black cells indicate the tasks that were not enacted.

The bilingual tasks with English (EN) as target were chosen by six different groups. On the contrary, English as source language did not receive much attention. French (FR) as target registered the highest number of submissions, but they were returned by a single participating team. Five Spanish groups participated in the

monolingual Spanish (ES) task, while in 2003 only the University of Alicante managed to run its system. New Dutch (NL) and Italian (IT) research groups registered in 2004 (only one Dutch group actually participated) in the corresponding monolingual tasks, which testifies the growing interest in QA for languages other than English. German (DE), that in 2003 was source language only, was chosen by two groups as target, like Portuguese (PT), at its first time at CLEF.

5 Results

Participants were allowed to submit just one response per question and up to two runs per task. Submissions were manually judged by human assessors, who considered both correctness and exactness of each answer.

A response was judged as correct when its form was clear and its content was responsive, while exactness is more related to the quantity than to the quality of the information retrieved by the systems. In the track guidelines [2], articles and prepositions were tentatively indicated as acceptable parts of speech that would not penalise the exactness of an answer. Adjectives, verbs and adverbs could instead add irrelevant or unnecessary information, as in the answer *Ex IMF Secretary General Dies* (that was returned in response to the question *Of what organisation was Pierre-Paul Schweitzer general manager?*), where only *IMF* would have been the exact and required string. At any rate, exactness was never precisely defined, so a certain degree of subjectivity in the judgements could not be eliminated.

In 2003, in order to facilitate participation, both exact and 50 bytes-long answer-strings were accepted (though assessed separately), but most participants chose to return exact responses. So, in 2004 only exact answers were allowed, which made the tasks more difficult. Responses were judged either as right, wrong, inexact or unsupported (when the answer-string contained a correct answer but the returned docid did not support it).

Factoid questions with the answer type MANNER (i.e. How- questions) and definition questions, that were included in the test sets in 2004 for the first time, needed more heuristically oriented evaluation criteria because their answers could be also long circumlocution or even entire sentences. In particular, answers to definition questions were judged considering their usefulness for a potential user that was assumed to know nothing of the person or the organisation addressed by the question. For instance, a correct answer returned in response to the question *Who is Jorge Amado?* was the following sentence: *American authors such as Stephen King and Sidney Sheldon are perennial best sellers in Latin American countries, while Brazilian Jorge Amado, Colombian Gabriel Garcia Marquez and Mexican Carlos Fuentes are renowned in U.S. literary circles.* In fact, it is clear from the sentence that Jorge Amado is a Brazilian writer and, moreover, it would have been difficult to extract a shorter and responsive string from this snippet.

The assessors were basically less demanding in terms of exactness when they judged these types of questions. However, accepting such long answers might be seen as equivalent to considering passage extraction rather than QA, so some judges disagreed on this subject. Because of the unnecessary information included in the answer-string above, some assessors would judge the response as inexact. No specific

assessment training was offered to all the groups, which should be taken into account in the future.

The organising group that had generated the questions in a particular language was in charge of the assessment of the runs with the same target language (except for the judgement of the English runs, that was taken over by the University of Limerick). As a common procedure, each run, containing 200 answers, was judged by more than one assessor. The DLTG group used a different approach, as described in section 5.2. The main measure was *accuracy*, that is the fraction of right answers. Answers had to be unranked (i.e. in the same order as in the test set), but a confidence value could be given for each response. Though it was not mandatory, this absolute value that could range between 0 and 1 was considered to calculate an additional C*onfidence-weighted Score* (CWS), borrowed from the TREC-2002 track [7]. Both accuracy and CWS reward systems for recognising correct answers, and both penalise them for mistaking wrong responses for correct ones. However, only CWS rewards systems that can predict their own performance.

The restriction to a single exact answer per question made the task harder than that proposed in 2003, when three ranked responses were accepted and the *Mean Reciprocal Rank* was computed. At CLEF 2003 the average performance was 41% of correct answers in the monolingual tasks and 25% in the cross-language ones, but if we consider just the first response to each question, the results drop to 29% and 17% respectively. In 2004 the average accuracy over the 20 runs submitted in the monolingual tasks was 23.7%, and 14.7% over the 28 bilingual runs. So, the average results of the two evaluation exercises are not so different, and the slight downgrade registered in 2004 is probably due to the introduction of the definition questions.

In the following seven sections the results of the runs for each target language are thoroughly discussed. For each target language two kinds of results are given in two separate tables. In the first one the systems' performance is described considering the number of right (R), wrong (W), inexact (X) and unsupported (U) answers that were returned, the overall accuracy, the partial accuracy on factoid and definition questions, the accuracy in recognising NIL questions (both Precision and Recall are given) and the Confidence-weighted Score of all the submitted runs. In the second table systems' accuracy is analysed with respect to the answer types of the questions in test set. Answer types are designated by the following abbreviations: *loc* ≡ LOCATION, *mea* ≡ MEASURE, *org* ≡ ORGANISATION, *per* ≡ PERSON, *man* ≡ MANNER, *obj* ≡ OBJECT, *oth* ≡ OTHER and *tim* ≡ TIME. Below each answer type, the number of posed questions of that type is shown in square brackets. The last row of the second table shows a virtual run, called *combination*, in which an answer is classified as right if any of the participating systems found it. This virtual run aims at showing the potential achievement if one merged all answers and considered the set of right answers, provided at least one answer per question were right.

5.1 Dutch as Target

Two research groups registered for tasks with Dutch as the target language, but only one team submitted runs: the University of Amsterdam, who had also participated in

2003. They submitted two monolingual runs, and one bilingual run (English to Dutch).

The Dutch test set contains 200 questions. Table 2 below details the results of the three submitted runs. Interestingly, on definition questions the bilingual English to Dutch run performed better than either of the two monolingual runs.

Table 2. Results of the monolingual and bilingual Dutch runs

Run Name	R (#)	W (#)	X (#)	U (#)	Overall Accuracy (%)	Accuracy over F (%)	Accuracy over D (%)	NIL Accuracy		CWS
								P	R	
uams041nlnl	88	98	10	4	44.00	42.37	56.52	0.00	0.00	-
uams042nlnl	91	97	10	2	45.50	45.20	47.83	0.56	0.25	-
uams041ennl	70	122	7	1	35.00	31.07	65.22	0.00	0.00	-

Table 3. Results of the Dutch runs, according to answer types of questions

Run Name	Given correct answers											
	Definition (#)		Factoid (#)								Total	
	org [11]	per [12]	loc [32]	man [15]	mea [15]	obj [10]	org [22]	oth [17]	per [49]	tim [17]	# [200]	%
uams041nlnl	6	7	14	3	6	1	10	5	26	10	88	44.00
uams042nlnl	4	7	15	3	4	1	11	5	30	11	91	45.50
uams041ennl	6	9	11	0	4	1	8	1	21	9	70	35.00
combination	7	10	20	3	8	2	13	5	36	16	120	60.00

The aim of the virtual run called *combination* is to provide an upper bound on the possible performance of a system that would merge the existing runs and somehow select the right answers from the combined pool of candidate answers. As an aside, this is actually how the University of Amsterdam's QA system works: separate streams each generate result files, and these are combined into a joint pool of candidate answers from which the final answers are selected.

5.2 English as Target

The work of assessing questions with English answers was assigned to the Documents and Linguistic Technology Group at Limerick. The five tasks enacted involved questions in Bulgarian, Finnish, French, German, Italian with English answers being returned from the *LA Times* (American English) *and Glasgow Herald* (Scottish English) collections. The starting point in carrying out the assessment comprised the TREC Evaluation Software written by Ellen Voorhees and the *Multieight-04* collection of manually retrieved answers.

Having studied the TREC software it was decided that it should be used on a question-by-question basis rather than on a run-by-run basis. This means that a single assessor reviews and evaluates all candidate answers to a given question. before moving to the next question. Originally we had envisaged that a given evaluator would assess all answers to different questions comprising a complete run before moving on to the next run.

The method used in carrying out the assessment was as follows. There were four primary assessors plus one secondary assessor. Each primary assessor - a native speaker of English - was assigned a set of questions, 1-50, 51-100, 101-150 and 151-200 respectively. The assessors, provided with a set of guidelines, then carried out their work, noting any doubtful cases. A series of meetings then took place at which these cases were considered in turn by all five assessors and a joint decision was made. To ensure consistency, the consequences of each decision were then cross-checked by each assessor against judgements of comparable cases. It should be noted therefore that while all responses to a particular question were judged by the same person, we did not use double-blind assessment where each judgement is made independently by two assessors.

Table 4. Results of the runs with English as target language

Run Name	R (#)	W (#)	X (#)	U (#)	Overall Accuracy (%)	Accuracy over F (%)	Accuracy over D (%)	NIL Accuracy		CWS
								P	R	
bgas041bgen	26	168	5	1	13.00	11.67	25.00	0.13	0.40	0.056
dfki041deen	47	151	0	2	23.50	23.89	20.00	0.10	0.75	0.177
dltg041fren	38	155	7	0	19.00	17.78	30.00	0.17	0.55	-
dltg042fren	29	164	7	0	14.50	12.78	30.00	0.14	0.45	-
edin041deen	28	166	5	1	14.00	13.33	20.00	0.14	0.35	0.049
edin041fren	33	161	6	0	16.50	17.78	5.00	0.15	0.55	0.056
edin042deen	34	159	7	0	17.00	16.11	25.00	0.14	0.35	0.052
edin042fren	40	153	7	0	20.00	20.56	15.00	0.15	0.55	0.058
hels041fien	21	171	1	0	10.88	11.56	5.00	0.10	0.85	0.046
irst041iten	45	146	6	3	22.50	22.22	25.00	0.24	0.30	0.121
irst042iten	35	158	5	2	17.50	16.67	25.00	0.24	0.30	0.075
lire041fren	22	172	6	0	11.00	10.00	20.00	0.05	0.05	0.032
lire042fren	39	155	6	0	19.50	20.00	15.00	0.00	0.00	0.075

We should point out that our reasoning and judgements were made with respect to the English versions of the questions. However, all the systems in this task group were using the 'same' questions in languages other than English. It is possible therefore that a question inadvertently asked something different in a particular language due to differences of translation. This could affect the results though perhaps not to a major degree.

Table 5. Results of the bilingual English runs, according to answer types of questions

Run Name	Given correct answers											
	Definition (#)		Factoid (#)								Total	
	org [11]	per [9]	loc [28]	man [15]	mea [20]	obj [12]	org [20]	oth [27]	per [28]	tim [30]	# [200]	%
bgas041bgen	2	3	5	2	1	2	1	2	4	4	26	13.00
dfki041deen	4	0	10	2	2	1	5	5	6	12	47	23.50
dltg041fren	3	3	8	5	2	1	1	2	4	9	38	19.00
dltg042fren	3	3	4	3	1	1	1	2	3	8	29	14.50
edin041deen	1	3	6	2	0	0	2	2	4	8	28	14.00
edin041fren	0	1	7	3	1	1	2	4	3	11	33	16.50
edin042deen	1	4	6	4	1	2	2	5	3	6	34	17.00
edin042fren	0	3	7	4	3	1	2	4	4	12	40	20.00
hels041fien[11]	0	0	3	0	2	0	5	4	5	2	21	10.88
irst041iten	0	5	11	0	1	0	6	3	8	11	45	22.50
irst042iten	0	5	5	0	1	0	2	5	6	11	35	17.50
lire041fren	3	1	9	0	1	0	3	0	1	4	22	11.00
lire042fren	2	1	13	0	1	0	4	1	6	11	39	19.50
combination	7	5	26	6	7	5	18	10	22	24	130	65.00

In Table 5 the results are sorted by category of questions. Some answer types (i.e. manner, measure and object) turned out to be difficult for systems, while the performance on location, factoid-person and time is quite good.

In making judgements concerning definitions we decided to err on the side of generosity and made no correction for the length of submissions although in practice these tended to be short. A response was considered correct if it provided salient information concerning the topic. Generally the task specification for such questions was considered somewhat vague and so the results while being interesting are not necessarily that informative. What seems to be necessary is a means of punishing answers which contain both relevant and irrelevant information. This has been attempted in TREC with mixed results. While the level of participation in the English target task group was very encouraging, the numbers participating was still very small in statistical terms and also varied from language pair to language pair. Therefore we should be careful not to conclude too much from the results in terms for example of the relative difficulty of different language pairs.

5.3 French as Target

A single research group took part in evaluation tasks with French as a target language: Neuchatel University. It took part in both monolingual and bilingual tasks. This participating team submitted 16 runs, two runs per source language, taken from the 8 available source languages: Bulgarian, German, English, Spanish, French, Italian,

[11] Since some typos were found in the FI=>EN test set, seven questions were not taken into consideration in the evaluation. None of them had received a right answer, so their exclusion did not affect the data in Table 5.

Dutch and Portuguese. In particular, two runs were submitted for the monolingual task.
Table 6 shows the assessment of the sixteen submitted runs. The monolingual runs appear in italics.

Table 6. Results of the monolingual and bilingual French runs

Run Name	R (#)	W (#)	X (#)	U (#)	Overall Accuracy (%)	Accuracy over F (%)	Accuracy over D (%)	NIL Accuracy		CWS
								P	R	
gine041bgfr	13	182	5	0	6.50	6.67	5.00	0.10	0.50	0.051
gine041defr	29	161	10	0	14.50	14.44	15.00	0.15	0.20	0.079
gine041enfr	18	170	12	0	9.00	8.89	10.00	0.05	0.10	0.033
gine041esfr	27	165	8	0	13.50	14.44	5.00	0.12	0.15	0.056
gine041frfr	*27*	*160*	*13*	*0*	*13.50*	*13.89*	*10.00*	*0.00*	*0.00*	*0.048*
gine041itfr	25	165	10	0	12.50	13.33	5.00	0.15	0.30	0.049
gine041nlfr	20	169	11	0	10.00	10.00	10.00	0.12	0.20	0.044
gine041ptfr	25	169	6	0	12.50	12.22	15.00	0.11	0.15	0.044
gine042bgfr	13	180	7	0	6.50	6.11	10.00	0.10	0.35	0.038
gine042defr	34	154	12	0	17.00	15.56	30.00	0.23	0.20	0.097
gine042enfr	27	164	9	0	13.50	12.22	25.00	0.06	0.10	0.051
gine042esfr	34	162	4	0	17.00	17.22	15.00	0.11	0.10	0.075
gine042frfr	*49*	*145*	*6*	*0*	*24.50*	*23.89*	*30.00*	*0.09*	*0.05*	*0.114*
gine042itfr	29	164	7	0	14.50	15.56	5.00	0.14	0.30	0.054
gine042nlfr	29	156	15	0	14.50	13.33	25.00	0.14	0.20	0.065
gine042ptfr	29	164	7	0	14.50	13.33	25.00	0.10	0.15	0.056

The best results were obtained for one of the monolingual runs (gine042frfr). This proves once again that it is *a priori* easier for the systems to answer correctly when the source language is the same as the target language. However, it is noticeable that the 2nd and 3rd best results are obtained by the two German-French runs (better than the other monolingual French run).
It is important to notice that the number of unsupported answers is 0 for all runs. This is expectable as all 16 runs are versions of the same system, and indicates that this system always supports the answers it gives.

The correct answers given for all the runs are presented in Table 7, clustered by answer type of questions.
Neuchatel system's weaknesses obviously lie in definition-organisation (recall 0%) and in factoid-manner (max. recall 21%) questions, whereas it gives its better results for definition-person (max. recall 50%), measure (32%) and location (34.5%) questions.

The virtual run in the last row, called combination, aims at getting an idea of what could be the expected potential performance of a system giving all the correct answers. The best run (gine042frfr) is able to supply only 50.51% of the correct

answers of "combination". This ratio could be enhanced if results for definition-organisation and factoid-manner, in particular, would be improved.

Table 7. Results of the monolingual and bilingual French runs, according to answer types of questions

Run Name	Given correct answers											
	Definition (#)		Factoid (#)								Total	
	org [8]	per [12]	loc [29]	man [14]	mea [28]	obj [15]	org [20]	oth [21]	per [32]	tim [21]	# [200]	%
gine041bgfr	0	1	1	3	2	2	1	1	2	0	13	6.50
gine041defr	0	3	6	0	5	3	4	2	4	2	29	14.50
gine041enfr	0	2	5	0	4	1	0	1	3	2	18	9.00
gine041esfr	0	1	7	0	4	3	3	2	4	3	27	13.50
gine041frfr	*0*	*2*	*8*	*0*	*8*	*0*	*1*	*3*	*2*	*3*	*27*	*13.50*
gine041itfr	0	1	3	1	5	3	4	2	3	3	25	12.50
gine041nlfr	0	2	6	1	5	1	1	2	1	1	20	10.00
gine041ptfr	0	3	5	0	5	2	1	2	3	4	25	12.50
gine042bgfr	0	2	2	1	2	2	0	2	2	0	13	6.50
gine042defr	0	6	7	0	5	3	3	2	6	2	34	17.00
gine042enfr	0	5	7	0	5	1	2	1	4	2	27	13.50
gine042esfr	0	3	8	0	4	2	5	3	4	5	34	17.00
gine042frfr	*0*	*6*	*10*	*0*	*9*	*1*	*6*	*6*	*4*	*7*	*49*	*24.50*
gine042itfr	0	1	5	1	4	3	4	3	4	4	29	14.50
gine042nlfr	0	5	5	0	7	2	2	4	3	1	29	14.50
gine042ptfr	0	5	5	0	5	2	2	3	3	4	29	14.50
combination	0	7	19	3	17	5	8	8	11	9	97	48.50

5.4 German as Target

Two research groups took part in tasks with German as target language, and only in the monolingual German task: DFKI, which had participated at CLEF-2003, and Fernuniversität Hagen, at its first participation, submitted one run each.

The German test set contained 200 questions. However, three questions contained spelling errors and were subsequently excluded from the evaluation, so that only 197 questions were taken into consideration.

Table 8. Results of the monolingual German runs

Run Name	R (#)	W (#)	X (#)	U (#)	Overall Accuracy (%)	Accuracy over F (%)	Accuracy over D (%)	NIL Accuracy		CWS
								P	R	
fuha041dede	67	128	2	0	34.01	31.64	55.00	0.14	1.00	0.333
dfki041dede	50	143	1	3	25.38	28.25	0.00	0.14	0.85	-

Table 8 shows the assessment of the two runs which were submitted. DFKI did not handle any definition questions. Both groups produced short and exact answers; no answer was longer than 6 words or 48 characters.

Table 9. Results of the monolingual German runs, according to answer types of questions

Run Name	Given correct answers											
	Definition (#)		Factoid (#)								Total	
	org [11]	per [9]	loc [22]	man [20]	mea [21]	obj [23]	org [23]	oth [22]	per [23]	tim [23]	# [197]	%
fuha041dede	6	5	12	4	4	4	5	7	10	10	67	34.01
dfki041dede	0	0	8	2	4	2	8	4	9	13	50	25.38
combination	6	5	14	4	5	4	11	8	13	16	86	43.65

The combination run in the last row shows that the best performing system (fuha041dede) is able to respond correctly to 78% of the questions that have been correctly answered by both teams in conjunction.

The DFKI group conducted an experiment to compare the QA system performance against human QA performance under time constraints [3]. Three subjects answered all 200 questions of the monolingual German test set with the help of a search engine. The time between the presentation of each question and the submission of the document ID was measured, and the answers were assessed. Only answers that were found within a given time limit were considered. Then the accuracy a human could achieve was calculated. It was found that a human who is allowed a maximum of 42 seconds per question achieves the same level of accuracy as the German "combination" run (DFKI run ≈ 30s, FUHA run ≈ 34s). In addition, the experiment revealed the difficulty of different answer types for humans, e.g., the average definition questions required 39 seconds and the average factoid questions 81 seconds.

5.5 Italian as Target

Two research groups took part in tasks with Italian as target language, and precisely only in the monolingual Italian task: ITC-Irst, that had participated also at CLEF-2003, and the Institute for Computational Linguistics in Pisa[12], at its first participation.

In 2003 ITC-Irst submitted two runs, and the system answered correctly at the first rank to 37.5% and 41.5% of the questions respectively. The lower results achieved in 2004 with the same system demonstrate that the task was harder. Nevertheless, as Table 10 shows, the overall accuracy of the runs ILCP and irst041 is over the average performance of the participants in the monolingual tasks.

[12] Joint work with the Department of Information and Communication Technology of the University of Pisa.

Table 10. Results of the monolingual Italian runs

Run Name	R (#)	W (#)	X (#)	U (#)	Overall Accuracy (%)	Accuracy over F (%)	Accuracy over D (%)	NIL Accuracy		CWS
								P	R	
ilcp041itit	51	117	29	3	25.50	22.78	50.00	0.62	0.50	-
irst041itit	56	131	11	2	28.00	26.67	40.00	0.27	0.30	0.155
irst042itit	44	147	9	0	22.00	20.00	40.00	0.66	0.20	0.107

The analysis of the results in Table 11 shows that *location*, *person* and *time* were the easiest answer types for the participating systems. How-questions constituted a problem for the Irst system, while ILCP answered four of them correctly, retrieving long text snippet that were judged as responsive. The accuracy over definition questions in all three submitted runs is relatively high. While the Irst system returned very short answers, trying to select the most relevant portion of text, ILCP system often gave long answer-strings, and many of them (14.5%) were judged as inexact, though they often contained the required information.
The runs ilcp and irst042 were the most precise in the whole track in identifying the questions with no response, though their recall is not so high.

Table 11. Results of the monolingual Italian runs, according to answer types of questions

Run Name	Given correct answers											
	Definition (#)		Factoid (#)								Total	
	org [11]	per [9]	loc [25]	man [12]	mea [30]	obj [10]	org [17]	oth [33]	per [28]	tim [25]	# [200]	%
ilcp041itit	5	5	9	4	3	2	2	4	5	12	51	25.50
irst041itit	5	3	8	1	6	3	5	3	8	14	56	28.00
irst042itit	5	3	7	0	3	2	2	2	8	12	44	22.00
combination	8	7	12	4	8	4	7	6	13	19	88	44.00

5.6 Portuguese as Target

Two research groups took part in tasks with Portuguese as target language, both in the monolingual task; one of them submitted two runs. None provided a confidence score. Since there was a duplicated question, (*Who was the first President of the United States?*), only 199 questions were taken into account in the summary statistics.

Table 12. Results of the monolingual Portuguese runs

Run Name	R (#)	W (#)	X (#)	U (#)	Overall Accuracy (%)	Accuracy over F (%)	Accuracy over D (%)	NIL Accuracy		CWS
								P	R	
ptue041ptpt	57	125	18	0	28.64	29.17	25.81	0.14	0.90	-
sfnx041ptpt	22	166	8	4	11.06	11.90	6.45	0.13	0.75	-
sfnx042ptpt	30	155	10	5	15.08	16.07	9.68	0.16	0.55	-

The table above shows the assessment of the three submitted runs. While the answers of the SFNX system were generally rather short, the PTUE system occasionally submitted longer answers (in one case, reaching 35 words).

Table 13. Results of the monolingual Portuguese runs, according to answer types of questions

Run Name	Given correct answers											
	Definition (#)		Factoid (#)								Total	
	org [14]	per [17]	loc [43]	man [4]	mea [23]	obj [6]	org [12]	oth [21]	per [44]	tim [15]	# [199]	%
ptue041ptpt	3	5	19	1	5	1	4	3	14	2	57	28.64
sfnx041ptpt	0	2	4	0	3	1	2	3	7	0	22	11.06
sfnx042ptpt	1	2	8	0	4	2	2	4	7	0	30	15.08
combination	3	6	25	1	5	3	4	6	19	2	74	37.18

5.7 Spanish as Target

Five groups submitted eight runs having Spanish both as target and source language. The test set contained 200 questions with the type distribution shown in Table 15.

Table 14. Results of the monolingual Spanish runs

Run Name	R (#)	W (#)	X (#)	U (#)	Overall Accuracy (%)	Accuracy over F (%)	Accuracy over D (%)	NIL Accuracy		CWS
								P	R	
aliv041eses	63	130	5	2	31.50	30.56	40.00	0.17	0.35	0.121
aliv042eses	65	129	4	2	32.50	31.11	45.00	0.17	0.35	0.144
cole041eses	22	178	0	0	11.00	11.67	5.00	0.10	1.00	-
inao041eses	45	145	5	5	22.50	19.44	50.00	0.19	0.50	-
inao042eses	37	152	6	5	18.50	17.78	25.00	0.21	0.50	-
mira041eses	18	174	7	1	9.00	10.00	0.00	0.14	0.55	-
talp041eses	48	150	1	1	24.00	18.89	70.00	0.19	0.50	0.087
talp042eses	52	143	3	2	26.00	21.11	70.00	0.20	0.55	0.102

Since, as Table 15 shows, some systems performed better for certain types of questions, the following question arises: why do we not reward specialisation? This issue has been explored in the Pilot Question Answering Task [4], in which the confidence score has been taken into account in the evaluation measure in order to reward systems' self-knowledge and answer validation when responding to different types of questions.

As the virtual *combination* run in the last row of Table 15 shows, the best performing system (aliv042eses) is able to respond correctly to only 57.5% of the questions that would have been correctly answered by all teams in conjunction. Systems show better behaviour when answering about locations, organisations, dates and persons. It is interesting to remark that, whereas individual systems show

important differences among the number of correct answers depending on the type of question, the combination of systems shows a quite uniform distribution.

Table 15. Results of the monolingual Spanish runs, according to answer types of questions

Run Name	Given correct answers											
	Definition (#)		Factoid (#)								Total	
	org [10]	per [10]	loc [22]	man [22]	mea [23]	obj [22]	org [23]	oth [22]	per [23]	tim [23]	# [200]	%
aliv042eses	7	2	6	4	6	4	12	6	7	11	65	32.50
aliv041eses	7	1	5	4	7	4	12	6	6	11	63	31.50
talp042eses	7	7	10	3	3	6	3	1	9	3	52	26.00
talp041eses	7	7	9	4	1	5	3	0	5	7	48	24.00
inao041eses	4	6	9	3	2	2	5	3	3	8	45	22.50
inao042eses	4	1	9	3	2	2	5	2	2	8	37	18.50
cole041eses	1	0	2	2	2	2	3	3	3	4	22	11.00
mira041eses	0	0	3	2	4	2	2	1	2	2	18	9.00
combination	7	9	16	7	10	9	15	11	14	15	113	56.50

Though different questions and different text collections were used, the overall results obtained for monolingual Spanish in 2004 are better than those in the 2003 track. The best result obtained in last edition was 40% of questions with a correct answer. However, three answers per question were allowed in 2003: if we consider only the percentage of correct answers found at the first rank, that was 24.5% for the best system, it is outperformed by the run aliv042eses, submitted by the University of Alicante, that in 2004 reached an accuracy of 32.5%.

6 Remarks on Evaluation

The four judgements adopted by the assessors (right, wrong, inexact and unsupported) have been used at TREC for many years and seem to cover most of the possible answers of a real QA system. Even so, the evaluation of the runs submitted at CLEF shows that sometimes they are somehow simplistic, and that they do not enable assessors to grasp the responsiveness of all the answers.

In particular, as the disagreement between assessors has shown, exactness is really difficult to judge, considering also that it has never been defined with objective criteria. The tentative rules we tried to draft concerning the acceptable and the unacceptable parts of speech did not always match with the sensibility of the human assessors. Furthermore, some types of questions, such as How- questions and definitions, have relatively long strings as answers, and for the time being it would be too demanding to require essential and not redundant responses. Maybe we should consider going back to the retrieval of short, meaningful passages (similar to the optional *justifications* that could be attached to the answers at TREC 2002), possibly rewarding those systems that are able to return just the minimal piece of information.

Alternatively, the judgement *inexact* could be kept, but differentiated so as to distinguish between an incomplete answer and one that is too long.

In addition, the judgement *unsupported* could be considered independently from *right* and *wrong* because assessors came across wrong answers that were completely unrelated to the document indicated in the docid.

Finally, an additional heuristic judgement that quantifies the *usefulness* of a response could be introduced; in fact an answer can be either wrong or inexact, but at the same time a potential user could draw some partial information from it.

As far as the NIL questions are concerned, they were usually generated using proper names or keywords that did not appear in the document collection. This procedure needs to be reconsidered, because a simple IR system could trivially identify them, though in 2004 the NIL accuracy was not very high. If NIL questions addressed entities that actually appear in the corpus, the task would be more challenging and significant.

Confidence-weighted score, that was used at TREC 2002 [7], could not be calculated for all the runs because the confidence value was not mandatory. When computed, it seemed to reflect the overall accuracy, and it does not provide further insight in the systems' performance.

7 Conclusions

Thanks to the high number of proposed tasks and to a growing interest in Question Answering by the European research community, the QA@CLEF-2004 attracted more participants than the previous edition. In addition, the benchmark resources built within the framework of these evaluation exercises contribute to the development and tuning of systems, and can be reused as training resources.

The results of the 2004 track are not fully comparable to those achieved in 2003, in fact the two tasks were designed differently: nonetheless, the accuracy in answering specific questions, such as those that had *location* and *time* as answer types, was encouragingly high in all the seven target languages. The introduction of definition and How- questions made the task harder, and the assessors encountered some difficulties in defining and judging objectively the responsiveness and exactness of the responses. It seems that in assessing these particular questions, it would be reasonable to accept short text passages instead of exact answer-strings. Besides, the evaluation process as it was designed, i.e. split over different sites with multiple assessors, lacked uniformity and would need stricter, common guidelines that cover as much as real output cases. This should as much as possible be reconsidered for future campaigns.

The evaluation measures adopted in 2004 followed closely the TREC-2002 QA track, but since the assessors sometimes found the four judgements (right, wrong, inexact and unsupported) inadequate, some changes might be introduced in the next exercises, aimed for instance at rewarding the usefulness of responses for a potential user. However, coming up with a user model that is useful, satisfactory, and realistic is highly non-trivial.

Acknowledgements

The authors would like to thank Donna Harman for her valuable feedback and suggestions in designing the track, and Ellen Voorhees for providing the NIST software for the assessment of the submitted runs.
Gregor Erbach wishes to thank the German Federal Ministry of Education and Research (BMBF) through the project COLLATE II (01 IN C02), that supported the work.
Bernardo Magnini and Alessandro Vallin have been partially supported by the WEBFAQ project funded by the Autonomous Province of Trento.
Anselmo Peñas has been partially supported by the Spanish Government under projects TIC-2002-10597-E and R2D2-Syembra TIC-2003-07158-C04-02.
Maarten de Rijke was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 612-13-001, 365-20-005, 612.069.006, 612.000.106, 220-80-001, 612.000.207, 612.066.302, and 264-70-050.
Paulo Rocha was supported by the Portuguese Fundação para a Ciência e Tecnologia, through grant POSI/PLP/43931/2001.

References

1. Burger, J., Cardie, C., Chaudhri, V., Gaizauskas, R., Harabagiu, S., Israel, D., Jacquemin, C., Lin, C.-Y., Maiorano, S., Miller, G., Moldovan, D., Ogden, B., Prager, J., Riloff, E., Singhal, A., Shrihari, R., Strzalkowski, T., Voorhees and E., Weishedel, R.: Issues Tasks and Program Structures to Roadmap Research in Question & Answering (2001).
   URL: http://www-nlpir.nist.gov/projects/duc/papers/qa.Roadmap-paper_v2.doc
2. CLEF 2004 Question Answering Track Guidelines (2004)
   URL: http://clef-qa.itc.it/2004/guidelines.html
3. Erbach, G.: Evaluating Human Question Answering Performance under Time Constraints, (2004)
   URL: http://purl.org/net/gregor/pub/human-qa/
4. Herrera, J., Peñas, A. and Verdejo, F.: Question Answering Pilot Task at CLEF 2004. In this volume
5. Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Peñas, A., Peinado, V., Verdejo, F. and de Rijke, M.: The Multiple Language Question Answering Track at CLEF 2003. In: Peters, C., Braschler, M., Gonzalo, J. and Kluck, M., (eds), Results of the CLEF 2003 Evaluation Campaign. Lecture Notes in Computer Science, Vol. 3237. Springer-Verlag, Berlin Heidelberg New York (2004)
6. Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Peñas, A., Peinado, V., Verdejo, F. and de Rijke, M.: Creating the DISEQuA Corpus: a Test Set for Multilingual Question Answering. In: Peters, C., Braschler, M., Gonzalo, J. and Kluck, M., (eds), Results of the CLEF 2003 Evaluation Campaign. Lecture Notes in Computer Science, Vol. 3237. Springer-Verlag, Berlin Heidelberg New York (2004)
7. Voorhees, E. M.: Overview of the TREC 2002 Question Answering Track. In: Voorhees, E. M. and Buckland, L. P., (eds), Proceedings of the Eleventh Text Retrieval Conference (TREC 2002). NIST Special Publication 500-251, Washington DC (2002) 115-123
8. Voorhees, E. M.: Overview of the TREC 2003 Question Answering Track. In: Voorhees, E. M. and Buckland, L. P., (eds), Proceedings of the Twelfth Text Retrieval Conference (TREC 2003). NIST Special Publication 500-255, Washington DC (2003) 54-68