The Multiple Language Question Answering Track at CLEF 2003

Bernardo Magnini*, Simone Romagnoli*, Alessandro Vallin*

Jesús Herrera**, Anselmo Peñas**, Víctor Peinado**, Felisa Verdejo**

Maarten de Rijke***

* ITC-irst, Centro per la Ricerca Scientifica e Tecnologica
Via Sommarive, 38050 Povo (TN), Italy.
{magnini,romagnoli,vallin}@itc.it

** UNED, Spanish Distance Learning University, Dpto. Lenguajes y Sistemas Informaticos
Ciudad Universitaria, c./Juan del Rosal 16, 28040 Madrid, Spain.
{jesus.herrera,anselmo,victor,felisa}@lsi.uned.es

*** Language and Inference Technology Group, ILLC, University of Amsterdam
Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands.
mdr@science.uva.nl

Abstract. This paper reports on the pilot question answering track that was carried out within the CLEF initiative this year. The track was divided into monolingual and bilingual tasks: monolingual systems were evaluated within the frame of three non-English European languages, Dutch, Italian and Spanish, while in the cross-language tasks an English document collection constituted the target corpus for Italian, Spanish, Dutch, French and German queries. Participants were given 200 questions for each task, and were allowed to submit up to two runs per task with up to three responses (either exact answers or 50 bytes long strings) per question.

We give here an overview of the track: we report on each task and discuss the creation of the multilingual test sets and the participants' results.

1 Introduction

The question answering (QA) track at TREC-8 represented the first attempt to emphasize the importance and foster research on systems that could extract relevant and precise information rather than documents. Question answering systems are designed to find answers to open domain questions in a large collection of documents. QA development has acquired an important role among the scientific community because it entails research in both natural language processing and information retrieval (IR), putting the two disciplines in contact. Differently from the IR scenario, a QA system processes questions formulated into natural language (instead of keyword-based queries) and retrieves answers (instead of documents).

The past TREC conferences laid the foundations for a formalized and widely accepted evaluation methodology of QA systems, but the three tracks organized so far focused just on monolingual systems for the English language, which constitutes a drawback we tried to address. We were mainly interested in testing multilingual systems, and in particular to push the QA community into designing them. As the number of the participants and the results achieved by their systems show, we can argue that in the field of multilingual QA there is much work to do. Within the frame of planning and coordinating the research on question answering, outlined in Maybury's roadmap, multilingual QA has a pivotal role and should deserve much attention in the next years. Multilinguality represents a new area in QA research, and a challenging issue toward the development of more complex systems [8].

Multilinguality enables the user to pose a query in a language that is different from the language of the reference corpus. The cross-language perspective could be quite useful when the required information is not available in the user's language (as it often happens surfing the web) and in particular it fits for the cultural situation in Europe, where different languages co-exist and are in contact, although English has become a widespread and standardized means of communication. In a multilingual environment, QA systems and other natural language processing resources could even contribute to conserve endangered languages that are progressively losing importance and prestige, in the effort to ensure their survival, as in the case of the 'Te Kaitito' bilingual question answering system for English and Maori [4].

Our activity, and in particular the production of two multilingual test sets that constitute reusable resources, can be regarded as a valuable contribution to the development of such cross-language systems [2]. The evaluation of cross-language resources is the key issue of the CLEF initiative, so our question answering track could not be limited to the English language. On the contrary, we attempted to raise interest on other European languages, like Italian, Spanish, Dutch, German and French. The basic novelty in comparison with the past TREC QA campaigns was the introduction of bilingual tasks, in which non-English queries are processed to find responses in an English document collection.

2 QA at CLEF

Our pilot question answering track was structured in both monolingual and bilingual tasks. We organized three monolingual tasks for Dutch, Italian and Spanish, in which the questions, the corpus and the responses were in the same language. In contrast, in the cross-language tasks we had Italian, Spanish, Dutch, French or German queries that searched for answers in an English document collection. In output, the systems had to retrieve English answers.

2.1 Monolingual Tasks

Unlike previous TREC QA tracks, we focused on the evaluation and on the production of reusable resources for non-English QA systems. The monolingual tasks were designed for three different languages: Dutch, Italian and Spanish. For each language we generated 200 queries, 180 of which were completely shared between all the three tasks. Participants were given the questions and the corresponding monolingual corpus: the task consisted in returning automatically, i.e. with no manual intervention, a ranked list of [docid, answer] pairs per question such that the retrieved document supported the answer. Participants were given 200 questions for each task, and were allowed to submit up to two runs per task with up to three responses per query. They could return either exact answers or 50 bytes long strings that contained the answer, although they were not allowed to use both modalities within the same run. Following the TREC model, we formulated 20 questions that had no known answer in the corpora: systems indicated their belief that there was no answer in the document collection by returning "NIL" instead of the [docid, answer] pair.

The monolingual Italian question answering task was planned and carried out under the co-ordination of the Italian Centro per la Ricerca Scientifica e Tecnologica (ITC-irst), that was in charge for the supervision of the whole QA track. We could use the document collections released at CLEF 2002, made up of articles drawn from a newspaper (*La Stampa*) and a press agency (*SDA*) of the year 1994. The entire Italian target corpus was 200 Mb wide (about 27 millions words) and it was made available to registered participants at the end of last January, so that they could test their systems using the document collection well in advance.

The UNED NLP group (Spanish Distance Learning University), as Spanish member of the CLEF consortium, was in charge for the monolingual Spanish task. The collection we were allowed to employ was the one released at CLEF 2002, i.e. more than 200,000 news from *EFE Press Agency* of the year 1994.

The Language and Inference Technology Group at the University of Amsterdam took care of the monolingual Dutch task. The collection used was the CLEF 2002 Dutch collection, which consists of two full years of the *Algemeen Dagblad* and *NRC Handelsblad* newspapers, adding up to about 200,000 documents of 540 Mb.

2.2 Cross-Language Tasks

Our interest in developing QA systems for languages other than English was not the only achievement we pointed at: the great novelty introduced in the CLEF QA track was multilinguality, whose potentialities are currently out of the scope of the TREC competition. Cross-language QA systems are crucially important when the language of the query and the language of the document collection are different, and in multicultural situations such a possibility is far from being remote. Searching information in the World Wide Web for instance is often difficult because the document retrieved is in a language we cannot understand. In this sense the cross-language tasks we organized represent a good chance to push the QA community to design and evaluate multilingual systems.

The cross-language tasks consisted in searching an English corpus to find English responses to queries posed in a different language. The target document collection we used was a corpus made up of *Los Angeles Times* articles of the year 1994, that was the same employed in last year's CLEF campaign. We translated into five languages the original two hundred English questions we generated, so we were able to organize five different bilingual tasks: Dutch, French, German, Italian and Spanish. As in the monolingual tasks, participants had to process 200 questions (15 had no answer in the corpus) posed in one of the five languages and could choose to submit either exact answers or 50 bytes strings, without mixing them in the same run.

2.3 Participants

Eight groups took part in this pilot question answering track, and a total of seventeen runs were submitted, three using 50 bytes long strings as answers and the other fourteen, in compliance with last year's TREC conditions, returning exact answers. The fact that most participants chose to retrieve exact answers shows that many have made the transition from more or less long strings to precise responses.
Table 1 below shows the name of the participants, the task in which they participated and the filename of their runs. It is interesting to notice that all the participants except the DFKI group had already participated in some previous TREC QA campaigns.

Table 1. Participants in the CLEF Question Answering Track. Note that the fifth and sixth letters in the run names show whether the responses are exact answers (ex) or 50 bytes long strings (st).

GROUP	TASK	RUN NAME
DLSI-UA U. of Alicante, Spain	Monolingual Spanish	alicex031ms alicex032ms
UVA U. of Amsterdam, the Netherlands	Monolingual Dutch	uamsex031md uamsex032md
ITC-irst Trento, Italy	Monolingual Italian	irstex031mi irstst032mi
	Bilingual Italian	irstex031bi irstex032bi
ISI U. of Southern California, USA	Bilingual Spanish	isixex031bs isixex032bs
/	Bilingual Dutch	/
DFKI Saarbruecken, Germany	Bilingual German	dfkist031bg
CS-CMU Carnegie Mellon U., USA	Bilingual French	lumoex031bf lumoex032bf
DLTG U. of Limerick, Ireland	Bilingual French	dltgex031bf dltgex032bf
RALI U. of Montreal, Canada	Bilingual French	udemst031bf udemex032bf

Three teams took part in the monolingual tasks, submitting a total of six runs. We had only one participant in each language, which is quite disappointing because no comparison can be made between similar runs. Anyway, since the question set for all the monolingual tasks was the same (except the NIL questions), the monolingual runs can be compared to some extent. Four teams initially registered for the monolingual Italian task, but unfortunately only one, the ITC-irst group, actually participated. Similarly, only the University of Alicante took part in the monolingual Spanish task submitting two runs of exact answers, although three other groups expressed their intention of participation. As for the monolingual Dutch task, the University of Amsterdam with its two runs of exact answers was the only participant.

Six groups participated in the cross-language tasks, submitting eleven runs. The challenging novelty of the cross-language question answering attracted more participants than the monolingual tasks: the bilingual French task was chosen by three groups, while no one tested their system in the bilingual Dutch.

3 Test Sets

From a potential user's point of view, a question answering system should be able to process natural language queries and return precise and unambiguous responses, drawn from a large reference corpus. Thus, in every evaluation campaign like the one we conducted, a set of well formulated questions is required. Since they should reflect real requests posed by humans, such questions must sound spontaneous and realistic. On the other hand, they must be clear, simple and factoid, i.e. related to facts, events, physical situations, so that the answers can be retrieved without inference. All the necessary information to answer the questions must be straightforwardly available and consequently included in the document collection searched by the systems. For this reason no external knowledge of the world should be required and the queries should deal with practical, concrete matters, rather than with abstract notions, that depend on personal opinion or reasoning.

The creation of the question sets for both the tasks entailed much work in terms of queries selection and answers verification. In order to establish some common criteria of comparison between the several languages involved, we decided to provide the participants, independently from the language, with the same queries. Thus, we created two collections of two hundred questions each, translated into different languages: one for the monolingual tasks and the other one for the cross-language tasks. As a result, we put together two reusable linguistic resources that can be useful for the QA community but also for other NLP fields, such as Machine Translation. The test set for the monolingual tasks in particular represents a multilingual collection of queries with their answers in different corpora.

3.1 Gold Standard for the Monolingual Tasks

The benchmark collection of queries and responses for the Dutch, Italian and Spanish monolingual tasks was the result of a joint effort between the coordinators, who decided to share the test sets in the three languages. Our activity can be roughly divided into four steps:

1. *Production of a pool of 200 candidate questions with their answers in each language*. These queries were formulated on the basis of the topics released by CLEF for the retrieval tasks of the year 2000, 2001 and/or 2002. The CLEF topics, i.e. a set of concepts chosen with the aim of covering the main events occurred in the years 1994 and/or 1995, allowed us to pose questions independently from the document collection. In this way we avoided any influence in the contents and in the formulation of the queries. Questions were posed according to common guidelines: they had to be generally short and fact-based, unrelated to subjective opinions. They could not ask for definitions (i.e. "Who is Bill Clinton") and they had to have just one unique and unambiguous item as response, which means that we avoided questions asking for multiple items like those used in the TREC list task. Three groups of native speakers, one for each language, were involved in this work and searched the correct answers. A question has an answer in the reference corpus if a document contains the correct response without any inference implying knowledge outside the document itself.
2. *Selection of 150 questions from each monolingual set*. Since our aim was to build a test set of shared queries that would find answers in all the monolingual corpora, each group chose 150 questions from its candidate pool and translated them into English, thus a larger collection of 450 queries was put together. English constituted a sort of inter-language we used to shift from one language to another, but in this phase we were aware that there was the risk of changing unwarily the content of the questions during the translation. Each group chose its 150 questions taking into consideration that they would be processed by the other two, so the most general queries, that were likely to find a response in the other two corpora, were selected. Those that were too strictly related to the specific issues of a country were discarded.
3. *Processing of the shared questions*. Once we had collected a pool of 450 questions that had response in one of the corpora, we detected the duplicates and eliminated them. Quite surprisingly, we found thirteen couples of queries that had an identical meaning, although the formulation could be slightly different. Then each group translated back from English the 300 questions provided by the other coordinators and verified whether they had an answer in its corpus.
4. *Selection of the final 200 questions*. At this point, about 450 different questions had been formulated and translated into Dutch, English, Italian and Spanish. All of them had at least one answer in at least one language (other than English), and more than 200, obtained by merging the data of the second cross-verification, proved to have at least one answer in all the three monolingual document collections. Our goal was to provide the QA participants with 200 questions, including a small rate of NIL queries, i.e. questions that do not have any known answer in the corpus. We agreed that the 10% of the test set was a reasonable amount of NIL questions, that were first introduced in QA evaluation at TREC-10 (2001). So we selected 180 questions from those that had a response in all the three corpora, and each group completed its monolingual test set adding 20 NIL questions, that were necessarily different for each task. Taking into consideration seven general classes of questions, we tried to balance the final test set of 180 questions, that is composed of: 45 entries that ask for the name or role of a PERSON, 40 that pertain a LOCATION, 31 a MEASURE, 23 an ORGANISATION, 19 a DATE, 9 a concrete OBJECT, while 13, due to their vagueness, can be labeled with OTHER.

The result of the question development phase is a useful and reusable multilingual question set, whose entries are structured in a XML format, as shown in the example of figure 1. More details are given in the paper “Creating the DISEQuA Corpus” (in this book).

```
<qa cnt="1" type="DATE">
    <language val="ITA" original="TRUE">
        <question assessor="Ale-irst">
            Quando è avvenuta la riunificazione delle due Germanie?
        </question>
        <answer n="1" idx="SDA19941115.00073">
            nel 1989
        </answer>
    </language>
    <language val="SPA" original="FALSE">
        <question assessor="Anselmo-UNED">
            ¿Cuándo se produjo la reunificación de Alemania?
        </question>
        <answer n="1" idx="EFE19941108-04388">
            1989
        </answer>
        <answer n="2" idx="EFE19941108-04508">
                 1989
        </answer>
    </language>
    <language val="DUT" original="FALSE">
        <question assessor="LIT">
            Wanneer vond de Duitse hereniging plaats?
        </question>
        <answer n="1" idx="NH19940128-0161">
            in 1989
        </answer>
    </language>
    <language val="ENG" original="FALSE">
        <question assessor="">
            When did the reunification of East and West Germany take place?
        </question>
        <answer n="1" idx="-1">
            SEARCH[in 1989]
        </answer>
    </language>
</qa>
```

Fig. 1. Gold Standard format of a question for the monolingual tasks

3.2 Gold Standard for the Cross-Language Tasks

While in the monolingual tasks we had three different document collections and three sets of questions, all the bilingual tasks had one English target corpus. For this reason we generated 200 English queries and verified manually that each of them (except 15 NIL) had at least an answer. Then the questions were translated into each language. As in the monolingual test sets, translators were asked to be as faithful as possible to the original English version, in fact we were aware that every translation could be different from the source.

Because of organizational problems encountered shortly before the test set creation deadline, three Italian native speakers at ITC-irst had to take on the job, even though there was a high risk of inconsistencies that may have affected the quality of the question set as a resource.

Due to time constraints we could not compile a large pool of general questions independently from the corpus and then verify them. Instead, we chose an alternative approach: we randomly selected a document from the collection (while trying to select news with a worldwide importance, avoiding sections that deal with local politics or issues too strictly related to Los Angeles counties) and picked up a text snippet that was relevant, long and interesting enough to get a question out of it. For instance, from the following passage

> The government has banned foods containing intestine or thymus from calves because a new scientific study suggested that they might be contaminated with the infectious agent of bovine spongiform encephalopathy, commonly called "mad cow disease”.

we drew the question ‘What is another name for the “mad cow disease”?’.

Finally, we obtained a benchmark corpus in which each question appears in six languages, as the tag attribute <language val> in figure 2 shows:

```
<qa cnt="4" type="OTHER">
    <language val="ENG"   original="TRUE">
        <question assessor="Ale-irst">
            What is another name for the "mad cow disease"?
        </question>
        <answer n="1" idx="LA091194.0096">
            bovine spongiform encephalopathy
        </answer>
    </language>
    <language val="ITA"   original="FALSE">
        <question assessor="Ale-irst">
            Qual è un altro nome per la "malattia della mucca pazza"?
        </question>
        <answer n="1" idx="">
            SEARCH[bovine spongiform encephalopathy]
        </answer>
    </language>
    <language val="SPA"   original="FALSE">
        <question assessor="">
            ¿Qué otro nombre recibe la enfermedad de las vacas locas?
        </question>
        <answer n="1" idx="">
            SEARCH[bovine spongiform encephalopathy]
```

```
      </answer>
    </language>
    <language val="DUT"   original="FALSE">
      <question assessor="">
        Wat is een andere naam voor "gekke-koeienziekte"?
      </question>
      <answer n="1" idx="">
        SEARCH[bovine spongiform encephalopathy]
      </answer>
    </language>
    <language val="GER"   original="FALSE">
      <question assessor="">
        Was ist ein anderer Name für "Rinderwahnsinn"?
      </question>
      <answer n="1" idx="">
        SEARCH[bovine spongiform encephalopathy]
      </answer>
    </language>
    <language val="FRE"   original="FALSE">
      <question assessor="">
        Quel autre nom donne-t-on à la "maladie de la vache folle"?
      </question>
      <answer n="1" idx="">
        SEARCH[bovine spongiform encephalopathy]
      </answer>
    </language>
  </qa>
```

Fig. 2. Gold Standard format of a question for the bilingual tasks

4 Results

Participants had one week to process the questions. Since no manual intervention of any kind was allowed, we asked participants to freeze their systems before downloading the queries from our "QA @ CLEF" website.[1] Before the start of the evaluation exercise, we released detailed guidelines with the necessary information about the required format of the submissions. We also put online a checking routine with which participants could make sure that their responses were in compliance with that.

4.1 Response Format

Since we allowed to submit both exact answers and 50 bytes long strings, we could not evaluate these two formats together. For this reason, we divided our track into two

[1] http://clef-qa.itc.it

subtasks with separated evaluations. The required format of the answers in both subtasks was the same, but we decided to draw up two separate results.

Table 2 shows an example of a participant's submissions, where the first column indicates the question number, provided by the organizers, and the string in the second one represents the unique identifier for a system and a run: the last two characters in this case show that the task is the bilingual Italian, and the fifth and sixth characters give information about the kind of responses retrieved in this run, i.e. exact answers.

The third field in the response format was the answer rank, which was crucially important for the evaluation of the system accuracy. Participants had to return the questions in the same order in which they had been downloaded, i.e. unranked. On the contrary, they had to rank their responses by confidence, putting in the first place the surest answer.

The integer or floating point score number of the fourth column justified the answer ranking. This field was not compulsory, and the systems that had no scoring strategies could set the value to default 0 (zero).

The docid, i.e. the unique identifier of the document that supports the given answer, is placed in the fifth column. If the system maintained that there was no answer in the corpus or if it could not find one, the docid was replaced by the string "NIL".

The answer string had to be given in the last field of the response, that was left empty when the docid was substituted by "NIL".

Table 2. Examples of responses drawn from the first bilingual run submitted by ITC-irst

0001	irstex031bi	1	3253	LA011694-0094	Modern Art
0001	irstex031bi	2	1776	LA011694-0094	UCLA
0001	irstex031bi	3	1251	LA042294-0050	Cultural Center
0002	irstex031bi	1	9	NIL	
0003	irstex031bi	1	484	LA012594-0239	1991
0003	irstex031bi	2	106	LA012594-0239	Monday
0004	irstex031bi	1	154	LA072294-0071	Clark
0004	irstex031bi	2	117	LA072594-0055	Huber
0004	irstex031bi	3	110	LA072594-0055	Department

4.2 Judgments and Evaluation Measures

Each single answer was judged by human assessors, who assigned to each response a unique label: either right, or wrong, or unsupported or inexact. Assessors were told to judge the submissions from a potential user's point of view, because the evaluation should take into consideration the future portability of QA systems. They analyzed both the answers themselves and the context, i.e. the document that supported the answer, in which they appeared.

Answers were judged to be incorrect (W) when the answer-string did not contain the answer or when the answer was not responsive. In contrast, a response was considered to be correct (R) when the answer-string consisted of nothing more than the exact, minimal answer (or contained the correct answer within the 50 bytes long string) and when the document returned supported the response. Unsupported answers (U) were correct but it was impossible to infer that they were responsive from the retrieved document. Answers were judged as non-exact (X) when the answer was correct and supported by the document, but the answer string missed bits of the response or contained more than just the exact answer.

In addition, we outlined some common criteria to distinguish and properly evaluate exact answers. We outlined general rules to apply in several cases: as regards the date of specific events that ended in the past, both day and year are normally required (unless the question refers only to the year), but if the day cannot be retrieved, the year is normally sufficient. For instance, if a system answered the question “When did Napoleon die?” returning “5th May”, it would be judged as incorrect. On the other hand, both “May 5, 1821” and “1821” could be correct exact answers. Actually, no clear definitions of exact answer have been formalized, yet. Discussing the issue, we noticed that, generally speaking, articles and prepositions do not invalidate an "exact" answer. So, both "July, 9" and "on the 9th of July" are exact answers. Similarly, appositions should not represent a problem, as well. So for instance, "1957", "year 1957" and "in the year 1957" should be exact answers, though someone could object that (with dates) "year" is redundant. When a query asks for a measure, the unit of measure can be accepted, too. So, both "30" and "30 degrees" are exact.

Concerning NIL answers, they are correct if neither human assessors nor systems have found any answer before or after the assessment process. If there is an answer in the collection, NIL is evaluated as incorrect. A NIL answer means that the system *believes* that there is not an answer for that question in the collection. There is no way for systems to explicitly indicate that they do not know or cannot find the answer for a question.

In strict evaluation, only correct answers (R) scored points, while in lenient evaluation the unsupported responses (U) were considered to be correct, too. The score of each question was the reciprocal of the rank for the first answer to be judged correct, which means that each query could receive either 1, or 0, or 0.333, or 0.5 points, depending on the confidence ranking.

The basic evaluation measure was the Mean Reciprocal Rank (MRR), that represents the mean score over all questions. MRR takes into consideration both recall and precision of the systems’ performance, and can range between 0 (no correct responses) and 1 (all the 200 queries have a correct answer at position one). Figures 6 and 7 below summarize the QA track results and show that the systems achieved better results in the monolingual than in the bilingual tasks, where the drop in performance is possibly due to the cross-language step. The QA system developed by ITC-irst proved to be the most accurate among those that participated, and the mean reciprocal rank scored in the monolingual Italian using 50 bytes long strings as answers was the highest of the whole QA track.

Answer responsiveness and exactness were in the opinion of human assessors, whose judgment could be different, as in everyday life we have different criteria to determine whether a response is good or not. During the evaluation of most of the runs, two different assessors judged each single question (each question of the bilingual runs were judged by three NIST assessors) and in case of discrepancies, they discussed their opinion and tried to reach an agreement. Whenever they could not agree, another person took the final decision.

Table 3. Examples of judged responses drawn from the first bilingual run submitted by ITC-irst

<table>
<tr><th colspan="7">Questions and judged responses</th></tr>
<tr><td colspan="7"><i>What museum is directed by Henry Hopkins?</i></td></tr>
<tr><td>W</td><td>1</td><td>irstex031bi</td><td>1</td><td>3252</td><td>LA011694-0094</td><td>Modern Art</td></tr>
<tr><td>U</td><td>1</td><td>irstex031bi</td><td>2</td><td>1773</td><td>LA011694-0094</td><td>UCLA</td></tr>
<tr><td>X</td><td>1</td><td>irstex031bi</td><td>3</td><td>1253</td><td>LA042294-0050</td><td>Cultural Center</td></tr>
<tr><td colspan="7"><b>Comment</b>: The second answer was correct but the document retrieved was not relevant. The third response missed bits of the name, and was judged non-exact.</td></tr>
<tr><td colspan="7"><i>Where did the Purussaurus live before becoming extinct?</i></td></tr>
<tr><td>W</td><td>2</td><td>irstex031bi</td><td>1</td><td>9</td><td>NIL</td><td></td></tr>
<tr><td colspan="7"><b>Comment</b>: The system erroneously “believed” that the query had no answer in the corpus, or could not find one.</td></tr>
<tr><td colspan="7"><i>When did Shapour Bakhtiar die?</i></td></tr>
<tr><td>R</td><td>3</td><td>irstex031bi</td><td>1</td><td>484</td><td>LA012594-0239</td><td>1991</td></tr>
<tr><td>W</td><td>3</td><td>irstex031bi</td><td>2</td><td>106</td><td>LA012594-0239</td><td>Monday</td></tr>
<tr><td colspan="7"><b>Comment</b>: In the questions that asked for the date of an event, the year was often regarded as sufficient.</td></tr>
<tr><td colspan="7"><i>Who is John J. Famalaro accused of having killed?</i></td></tr>
<tr><td>W</td><td>4</td><td>irstex031bi</td><td>1</td><td>154</td><td>LA072294-0071</td><td>Clark</td></tr>
<tr><td>R</td><td>4</td><td>irstex031bi</td><td>2</td><td>117</td><td>LA072594-0055</td><td>Huber</td></tr>
<tr><td>W</td><td>4</td><td>irstex031bi</td><td>3</td><td>110</td><td>LA072594-0055</td><td>Department</td></tr>
<tr><td colspan="7"><b>Comment</b>: The second answer, that returned the victim’s last name, was considered sufficient and correct, since in the document retrieved no other people named “Huber” were mentioned.</td></tr>
</table>

After the submission deadline had passed, we detected some mistakes in the questions. In particular, a blunder persisted in the Italian queries: we wrongly put an apostrophe after the contraction of the question word “quale” (“which”/”what”). We found 21 cases in the monolingual test set and 17 cases in the bilingual one. In the TREC campaigns the questions that contain mistakes are excluded from the evaluation, but, considering that the form “qual’e’/era” is quite common in Italian and that a QA system should be robust enough to recognize variant spellings, we decided to keep those queries. For the sake of completeness, we calculated precision and recall without the questions with that mistake, and we obtained just a very minor variation of the values (around 1%).

Translation could be the source of mistakes, as well. In the monolingual Spanish questions collection, “minister of Foreign Affairs” was erroneously translated as "president of Foreign Affairs" during the question sharing between the Italian and the Spanish coordinators.

In tables 4 and 6 below, we give a general overview of the results achieved by participants. In the monolingual exact answers runs there was a certain homogeneity in the performance, in fact there was not a great gap between the average (81 questions answered correctly) and the best result (97 in strict evaluation).

Differently, the results of the bilingual exact answers runs show a clear drop in the systems' accuracy: the difference between the best result (90 queries with at least a right answer) and the average (51) seems to be significant.

Concerning the 50 bytes long answers runs (tables 5 and 7), they do not lend themselves to many interpretations: we allowed to submit also these longer responses to facilitate and attract as many participants as possible, but in the end just three groups decided to return them, so we cannot make significant comparisons. The TREC workshops have probably pushed the QA community in tuning the systems on exact answers, and actually, it seems that there is not a great difference between exact and 50 bytes answers. In next year's campaign we could keep both exact and longer answers, maybe expanding the latter ones to 200 bytes or more.

Table 4. Summary statistics of the exact answers runs

EXACT ANSWERS RUNS

	GROUP	TASK	RUN NAME	MRR		# of Q. with at least one right answer		NIL Answers	
				Str.	Len.	Str.	Len.	total	R
MONO-LINGUAL TASKS	DLSI-UA	Monoling. Spanish	alicex031ms	.307	.320	80	87	21	5
			alicex032ms	.296	.317	70	77	21	5
	ITC-irst	Monoling. Italian	irstex031mi	.422	.442	97	101	4	2
	UVA	Monoling. Dutch	uamsex031md	.298	.317	78	82	200	17
			uamsex032md	.305	.335	82	89	200	17
CROSS-LANGUAGE TASKS	ISI	Bilingual Spanish	isixex031bs	.302	.328	69	77	4	0
			isixex032bs	.271	.307	68	78	4	0
	ITC-irst	Bilingual Italian	irstex031bi	.322	.334	77	81	49	6
			irstex032bi	.393	.400	90	92	28	5
	CS-CMU	Bilingual French	lumoex031bf	.153	.170	38	42	92	8
			lumoex032bf	.131	.149	31	35	91	7
	DLTG	Bilingual French	dltgex031bf	.115	.120	23	24	119	10
			dltgex032bf	.110	.115	22	23	119	10
	RALI	Bilingual French	udemex032bf	.140	.160	38	42	3	1

Table 5. Number of correct answers at a given rank in the exact answers runs. As can be noticed, all the systems (except DLTG's one) returned more than one answer per question, and ranked the responses quite well (i.e. placing most of them at the first place).

RIGHT ANSWERS RANKING

RUN NAME	STRICT				LENIENT			
	1st	2nd	3rd	total	1st	2nd	3rd	total
alicex031ms	49	13	18	80 (40%)	49	15	23	87 (43.5%)
alicex032ms	51	12	7	70 (35%)	53	15	9	77 (38.5%)
irstex031mi	75	13	9	97 (48.5%)	79	13	9	101 (50.5%)
uamsex031md	47	14	17	78 (39%)	50	17	15	82 (41%)
uamsex032md	46	19	17	82 (41%)	50	25	14	89 (44.5%)
isixex031bs	53	13	3	69 (34.5%)	56	16	5	77 (38.5%)
isixex032bs	43	18	7	68 (34%)	48	21	9	78 (39%)
irstex031bi	55	13	9	77 (38.5%)	56	15	10	81 (40.5%)
irstex032bi	70	12	8	90 (45%)	71	13	8	92 (46%)
lumoex031bf	25	8	5	38 (19%)	28	9	5	42 (21%)
lumoex032bf	22	8	1	31 (15.5%)	25	9	1	35 (17.5%)
dltgex031bf	23	0	0	23 (11.5%)	24	0	0	24 (12%)
dltgex032bf	22	0	0	22 (11%)	23	0	0	23 (11.5%)
udemex032bf	20	12	6	38 (19%)	23	13	6	42 (21%)

Table 6. Summary statistics of the 50 bytes long answers runs

50 BYTES LONG ANSWERS RUNS

	GROUP	TASK	RUN NAME	MRR		# of Q. with at least one right answer		NIL Answers	
				Str.	Len.	Str.	Len.	total	R
MONO-LING.	**ITC-irst**	Monoling. Italian	irstst032mi	.449	.471	99	104	5	2
CROSS LANG.	**DFKI**	Bilingual German	dfkist031bg	.098	.103	29	30	18	0
	RALI	Bilingual French	udemst031bf	.213	.220	56	58	4	1

Table 7. Number of correct answers at a given rank in the 50 bytes long answers runs.

RIGHT ANSWERS RANKING

RUN NAME	STRICT				LENIENT			
	1st	2nd	3rd	total	1st	2nd	3rd	total
irstst032mi	83	9	7	99 (49.5%)	87	10	7	104 (52%)
dfkist031bg	13	8	8	29 (14.5%)	14	8	8	30 (15%)
udemst031bf	32	16	8	56 (28%)	33	17	8	58 (29%)

Tables 5 and 7 show that the systems were quite accurate in ranking their correct answers: in strict evaluation, about the 70% of the correct responses was returned at the first rank, on the average.

Strict and lenient evaluation results actually do not differ much. This suggests that the systems are quite precise in the correct answers they return: the unsupported responses were in fact a few. More strikingly, the performance of the cross-language systems turned out to be quite low, which suggests that multilinguality is a field that requires much more attention and investigation.

5 Conclusions and Future Perspectives

The first European evaluation of non-English QA systems has given rise to useful resources for future multilingual QA developments. It has allowed us to establish and test a methodology and criteria for both the test suit production and the assessment procedure. Unfortunately, the CLEF QA Track did not receive the expected attention in terms of participation, and in most tasks just one group submitted its results. Actually, twelve research groups registered and were interested into participating, but some of them could not adjust their system on time. This suggests that the debate and the activities on multilingual QA have a certain appeal on the community, even though much challenging work remains to be done. We can be pleased of the outcome of this pilot QA evaluation exercise, and we hope that the results and the resources we developed will encourage many other groups to participate in future campaigns.

Cross-linguality has always been out of the scope of the TREC QA tracks, and our pilot QA at CLEF hopefully represents a first step in the direction of more sophisticated evaluation campaigns of multilingual systems. In our track, we provided five non-English question sets but just one English target document collection: in the future we could have several reference corpora in different languages, many different question sets and answers translated into different languages. Multilinguality provides us with the opportunity to experiment with different approaches, exploring many potential applications: for instance, we could think about developing intelligent systems that taking into consideration the language and the text coverage, select the most useful target corpus to search the answer for a particular question posed in a particular language. The possibilities are manifold, and our cross-language tasks can be considered just a starting point.

6 Acknowledgements

The work described in this paper has been supported by the Autonomous Province of Trento, in the framework of the WebFAQ Project, by the Spanish Government (MCyT, TIC-2002-10597-E) and 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, and 612.000.207.

The authors wish to acknowledge the contribution and support given by Franca Rossi, Elisabetta Fauri, Pamela Forner and Manuela Speranza at ITC-irst, who helped us in the generation and verification of the questions for both the monolingual and bilingual tasks.

Stephan Busemann at DFKI and Jian-Yun Nie at the University of Montreal took on the job of translating into German and French the questions for the bilingual tasks, and their contribution was fundamental.

We wish to thank the NIST, and in particular Ellen Voorhees and Donna Harman, for judging all the bilingual runs and for providing us with the necessary resources and feedback we needed to organize this pilot track.

We are also grateful to Carol Peters (ISTI-CNR) and Charles Callaway (ITC-irst) who, as English native speakers, edited the questions for the cross-language tasks.

We wish to thank Henry Chinaski, Vera Hollink, and Valentin Jijkoun for their help in the development and assessment of the monolingual Dutch task.

References

1. Braschler, M. and Peters, C.: The CLEF Campaigns: Evaluation of Cross-Language Information Retrieval Systems. UPGRADE (The European Online Magazine for the IT Professional), Vol. III, Issue no. 3, (June 2002).
URL: http://www.upgrade-cepis.org/issues/2002/3/up3-3Braschler.pdf .
2. 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, E., Weishedel, R.: Issues Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A), 2001.
URL: http://www-nlpir.nist.gov/projects/duc/papers/qa.Roadmap-paper_v2.doc.
3. CLEF 2003 Question Answering Track: Guidelines for the Monolingual and Bilingual Tasks.
URL: http://clef-qa.itc.it/guidelines.htm.
4. Knott A., Bayard I., de Jager S., Smith L., Moorfield J. and O'Keefe R.: A Question-Answering System for English and Maōri. Proceedings of the Fifth Biannual Conference on Artificial Neural Networks and Expert Systems (ANNES), University of Otago, November 2001.

5. Liddy, E.D.: Why are People Asking these Questions? A Call for Bringing Situation into Question-Answering System Evaluation. LREC Workshop Proceedings on Question Answering – Strategy and Resources, Grand Canary Island, Spain, 2002.
6. Magnini B., Negri M., Prevete R., Tanev H.: Multilingual Question/Answering: the DIOGENE System. Proceedings of the Tenth Text REtrieval Conference (TREC-2001), Gaithersburg, MD., 2001.
7. Magnini, B.: Evaluation of Cross-Language Question Answering Systems, proposal presentation held at the CLEF Workshop 2002.
   URL: http://clef.iei.pi.cnr.it:2002/workshop2002/presentations/q-a.pdf.
8. Maybury , M.: Toward a Question Answering Roadmap, 2002.
   URL:www.mitre.org/work/tech_papers/tech_papers_02/maybury_toward
9. Voorhees, E. M.: The TREC-8 Question Answering Track Report. Proceedings of the Eighth Text REtrieval Conference (TREC-8), Gaithersburg, MD., 1999.
10. Voorhees, E. M.: Overview of the TREC 2001 Question Answering Track. Proceedings of the Tenth Text REtrieval Conference (TREC-2001), Gaithersburg, MD., 2001.
11. Voorhees, E. M.: Overview of the TREC 2002 Question Answering Track. Proceedings of the Eleventh Text REtrieval Conference (TREC-2002), Gaithersburg, MD., 2002.
12. Voorhees, E. M., Tice, D. M.: Building a Question Answering Test Collection. Proceedings of SIGIR2000, Athens, Greece, 2000.