Overview of the CLEF 2007 Multilingual Question Answering Track

Danilo Giampiccolo[1], Pamela Forner[1], Jesús Herrera[2], Anselmo Peñas[3], Christelle Ayache[4], Corina Forascu[5], Valentin Jijkoun[6], Petya Osenova[7], Paulo Rocha[8], Bogdan Sacaleanu[9], and Richard Sutcliffe[10]

[1] CELCT, Trento, Italy
{giampiccolo,forner}@celct.it
[2] Departamento de Ingeniería del Software e Inteligencia Artificial, Universidad Complutense de Madrid, Spain
jesus.herrera@fdi.ucm.es
[3] Departamento de Lenguajes y Sistemas Informáticos, UNED, Madrid, Spain
anselmo@lsi.uned.es
[4] ELDA/ELRA, Paris, France
ayache@elda.fr
[5] Faculty of Computer Science, University "Al. I. Cuza" of Iaşi, Romania Institute for Computer Science, Romanian Academy, Iaşi, Romania
corinfor@info.uaic.ro
[6] Informatics Institute, University of Amsterdam, The Netherlands
jijkoun@science.uva.nl
[7] BTB, Bulgaria
petya@bultreebank.org
[8] Linguateca, SINTEF ICT, Norway and Portugal
Paulo.Rocha@alfa.di.uminho.pt
[9] DFKI, Germany
Bogdan.Sacaleanu@dfki.de
[10] DLTG, University of Limerick, Ireland
richard.sutcliffe@ul.ie

Abstract. The fifth QA campaign at CLEF [1], having its first edition in 2003, offered not only a main task but an Answer Validation Exercise (AVE) [2], which continued last year's pilot, and a new pilot: the Question Answering on Speech Transcripts (QAST) [3, 15]. The main task was characterized by the focus on cross-linguality, while covering as many European languages as possible. As novelty, some QA pairs were grouped in clusters. Every cluster was characterized by a topic (not given to participants). The questions from a cluster possibly contain co-references between one of them and the others. Finally, the need for searching answers in web formats was satisfied by introducing Wikipedia[1] as document corpus. The results and the analyses reported by the participants suggest that the introduction of Wikipedia and the topic related questions led to a drop in systems' performance.

[1] http://wikipedia.org

C. Peters et al. (Eds.): CLEF 2007, LNCS 5152, pp. 200–236, 2008.

1 Introduction

Inspired in previous TREC evaluation campaigns, QA tracks have been proposed at CLEF since 2003. During these years, the effort of the organizers has been focused on two main issues. One of them was to offer an evaluation exercise characterized by cross-linguality, covering as many languages as possible. From this perspective, major attention has been given to European languages, not only adding at least one new language every year, but maintaining the catalogue of offered ones, except for Finish, which only could be offered in the 2005 edition. The other important issue was to maintain a balance between the established procedure inherited from the TREC campaigns and innovation. This allowed newcomers to join the competition and, at the same time, offered "veterans" more challenges. Following these principles, in QA@CLEF 2007 a pilot task on *Question Answering on Speech Transcripts* and a subsidiary task on Answer Validation (AVE) were proposed together with a *main* task. As far as the latter is concerned, the most significant innovations were the introduction of topic-related questions and the possibility to search for answers in Wikipedia. The topic-related questions consisted of clusters of questions which were related to the same topic. The requirement for related questions on a topic necessarily implies that the questions will refer to common concepts and entities within the domain in question. This accomplished either by co-reference either by anaphoric reference to the topic declared implicitly in the first question or in its answer. As far as the other major innovation of this year's campaign, beside the data collections composed of news articles provided by ELRA/ELDA, also Wikipedia was considered, capitalizing on the experience of the WiQA pilot task proposed in 2006.

As general remark, the positive trend in participation registered in the previous campaigns was inverted for first time in the history of the QA@CLEF.

As reflected in the results, the task proved to be more difficult than expected, as in comparison with last year's results dropped both in the multi-lingual subtasks and in the monolingual subtasks.

QA@CLEF 2007 was carried out according to the spirit of the campaign, consolidated in previous years. Beside the classical main task, an *Answer Validation Exercise* [13] and a pilot task on *Question Answering on Speech Transcripts* [15] were proposed:

- the *main* task, divided into several monolingual and bi-lingual sub-tasks, is described in this paper.
- the *Answer Validation Exercise* (AVE) continued the successful experiment proposed in 2006. In this task, systems were required to emulate human assessment of QA responses and decide whether an *Answer* to a *Question* is correct or not according to a given *Text*. Results were evaluated against the QA human assessments [2]. The overview of this exercise can be found in this volume [13].
- the *Question Answering on Speech Transcripts* (QAST) pilot task aimed at providing a framework in which QA systems can be evaluated when the answers to factual and definition questions must be extracted from spontaneous speech transcriptions. The main goals of this pilot were:
  - comparing the performances of the systems dealing with both types of transcriptions

- measuring the loss of each system due to the state of the art of the Automatic Speech Recognition (ASR) technology.
- in general, motivating and driving the design of novel and robust factual QA architectures for automatic speech transcriptions [3]. The overview of this exercise can be found in this volume [15].

This paper describes the preparation process and presents the results of the QA track at CLEF 2007. In section 2, the tasks of the track are described in detail. The results are reported in section 3. In section 4, some final analysis about this campaign is given. And section 5 consists of a draft about what should be addressed in the near future of QA@CLEF.

2 Task Description

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

Following the example of TREC, this year the exercise consisted of topic-related questions, i.e. clusters of questions which were related to the same topic and possibly contained co-references between one question and the others. Neither the question types (F, D, L) nor the topics were given to the participants.

The systems were fed with a set of 200 questions -which could concern facts or events (F-actoid questions), definitions of people, things or organisations (D-efinition questions), or lists of people, objects or data (L-ist questions)- and were asked to return one exact answer, where *exact* meant that neither more nor less than the information required was given.

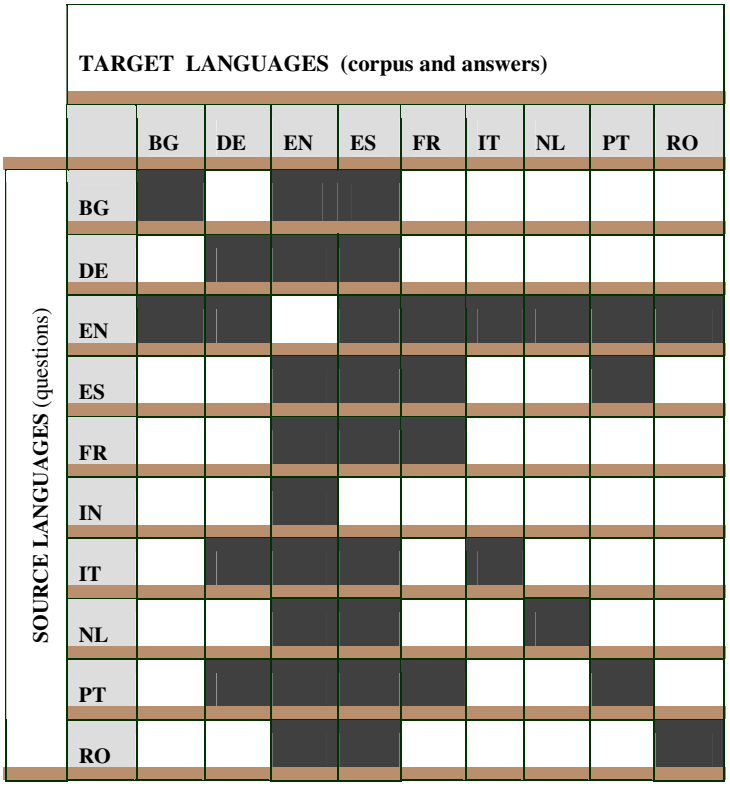
The answer needed to be supported by the docid of the document in which the exact answer was found, and by portion(s) of text, which provided enough context to support the correctness of the exact answer. Supporting texts could be taken from different sections of the relevant documents, and could sum up to a maximum of 700 bytes. There were no particular restrictions on the length of an answer-string, but unnecessary pieces of information were penalized, since the answer was marked as *ineXact*. As in previous years, the exact answer could be exactly copied and pasted from the document, even if it was grammatically incorrect (e.g.: inflectional case did not match the one required by the question). Anyway, systems were also allowed to use natural language generation in order to correct morpho-syntactical inconsistencies (e.g., in German, changing *dem Presidenten* into *der President* if the question implies that the answer is in nominative case), and to introduce grammatical and lexical changes (e.g., QUESTION*: What nationality is X*? TEXT: *X is from the Netherlands* → EXACT ANSWER: Dutch).

The subtasks were both:

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

Ten source languages were considered, namely, Bulgarian, Dutch, English, French, German, Indonesian, Italian, Portuguese, Romanian and Spanish. All these languages

Table 1. Tasks activated in 2007 (coloured cells)



were also considered as target languages, except for Indonesian, which had no news collections available for the queries and, as was done in the previous campaigns, used the English question set translated into Indonesian (IN).

As shown in Table 1, 37 tasks were proposed:

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

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

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

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

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

2.1 Questions Grouped by Topic

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

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

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

Q1: *Who is George W. Bush?*; Q2: *When was he born?*; Q3: *Who is his wife?*

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

Q4: *What is a polygraph?*; Q5: *When was **it** invented?*

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

Q6: *Who wrote the song "Dancing Queen"?*; Q7: *How many people were in **the group**?*

Here *the group* refers to an entity expressed not in the question but only in the answer. However the QA system does not know this and has to infer it, a task which can be very complex, especially if the topic is not provided in the test set.

2.2 Addition of Wikipedia

Another major innovation of this year's campaign concerned the corpora at which the questions were aimed. In fact, beside the data collections composed of news articles provided by ELRA/ELDA (see Table 3), also Wikipedia was considered, capitalizing on the experience of the WiQA pilot task proposed in 2006 [9].

The Wikipedia pages in the target languages, as found in the version of November 2006, could be used. Romanian, which was addressed as a target language for the first

Table 3. Document collections used in QA@CLEF 2007

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

time, had Wikipedia[2] as the only document collection, because there was no newswire Romanian corpus. The "snapshots" of Wikipedia were made available for download both in XML and HTML versions. The answers to the questions had to be taken from actual entries or articles of Wikipedia pages. Other types of data such as images, discussions, categories, templates, revision histories, as well as any files with user information and meta-information pages, had to be excluded.

One of the major reasons for using Wikipedia was to make a first step towards web formatted corpora where to search for answers.

As nowadays so large information sources are available on the web, this is may be considered a desirable next level in the evolution of QA systems. An important advantage of Wikipedia is that it is freely available for all languages so far considered. Anyway the variation in size of Wikipedia, depending on the language, is still problematic.

[2] http://static.wikipedia.org/downloads/November_2006/ro/

2.3 Types of Questions

As far as the question types are concerned, as in previous campaigns, the three following categories were considered:

1. *Factoid questions*, fact-based questions, asking for the name of a person, a location, the extent of something, the day on which something happened, etc. We consider the following 8 answer types for factoids:
   - PERSON, e.g.: Q8: *Who was called the "Iron-Chancellor"?* A8: *Otto von Bismarck*.
   - TIME, e.g.: Q9: *What year was Martin Luther King murdered?* A9: *1968*.
   - LOCATION, e.g.: Q10: *Which town was Wolfgang Amadeus Mozart born in?* A10: *Salzburg*.
   - ORGANIZATION, e.g.: Q11: *What party does Tony Blair belong to?*: A11: *Labour Party.*
   - MEASURE, e.g.: Q12: *How high is Kanchenjunga?* A12: *8598m.*
   - COUNT, e.g.: Q13: *How many people died during the Terror of PoPot?* A13: *1 million.*
   - OBJECT, e.g.: Q14: *What does magma consist of?* A14: *Molten rock.*
   - OTHER, i.e. everything that does not fit into the other categories above, e.g.: Q15: *Which treaty was signed in 1979?* A15: *Israel-Egyptian peace treaty.*
2. *Definition questions*, questions such as "What/Who is X?", and are divided into the following subtypes:
   - PERSON, i.e., questions asking for the role/job/important information about someone, e.g.: Q16*: Who is Robert Altmann?* A16*: Film maker*
   - ORGANIZATION, i.e., questions asking for the mission/full name/important information about an organization, e.g.: Q17: *What is the Knesset?* A17: *Parliament of Israel.*
   - OBJECT, i.e., questions asking for the description/function of objects, e.g.: Q18: *What is Atlantis?* A18: *Space Shuttle.*
   - OTHER, i.e., question asking for the description of natural phenomena, technologies, legal procedures etc., e.g.: Q19: *What is Eurovision?* A19: *Song contest.*
3. *closed list questions:* i.e., questions that require one answer containing a determined number of items, e.g.: Q20: *Name all the airports in London, England.* A20: *Gatwick, Stansted, Heathrow, Luton and City.*

As only one answer was allowed, all the items had to be present in sequence in the document and copied, one next to the other, in the answer slot.

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

Q21: *Who was the Chancellor of Germany from 1974 to 1982?*
A21: *Helmut Schmidt.*

Q22: *Which book was published by George Orwell in 1945?*
A22: *Animal Farm.*

Q23: *Which organization did Shimon Perez chair after Isaac Rabin's death?*
A23: *Labour Party Central Committee.*

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

The distribution of the questions among these categories is described in Table 4. Each question set was then translated into English, which worked as inter-language during the translation of the datasets into the other tongues for the activated cross-lingual subtasks.

Table 4. Test set breakdown according to question type, number of participants and number of runs

	F	D	L	T	NIL	# Participants	# Runs
BG	158	32	10	12	0	0	0
DE	164	28	8	27	0	2	5
EN	161	30	9	3	0	5	8
ES	148	42	10	40	21	5	5
FR	148	42	10	40	20	2	2
IT	147	41	12	38	20	1	1
NL	147	40	13	30	20	0	0
PT	143	47	9	23	18	6	8
RO	160	30	10	52	7	1	2

2.4 Formats

As the format is concerned, this year both input and output files were formatted as an XML file. For example, the first three questions in the EN-DE test set, i.e. English questions that hit a German document collection - were represented as follows:

```
- <input>
 <q target="DE" source="EN" id="0001" group_id="3600">Who is the singer
    of the band U2?</q>
 <q target="DE" source="EN" id="0002" group_id="3601">What age did El-
    vis Presley die?</q>
 <q target="DE" source="EN" id="0003" group_id="3602">Where is the
    Statue of Liberty located?</q>
```

An example of system output which answered the above questions was the following:

```
 <?xml version="1.0" encoding="UTF-8" standalone="yes" ?>
- <output>
- <a score="1.4222687" run_id="dfki071ende" group_id="3600" q_id="0001">
  <answer>Annett Louisan</answer>
- <support>
  <s_id>de_xml/p01/161699.xml</s_id>
```

```
  <s_string>Aber auch Sänger und Sängerinnen wie Kate Ryan, Annett Loui-
     san, die Band Overground, die Band Sportfreunde Stiller, Xavier
     Naidoo, Die Fantastischen Vier, das DJ-Duo Blank &amp;amp; Jones und
     Melendiz nutzten die U-Bahn und Kate Ryan Bahnhöfe für Kate Ryan
     Musikvideos.</s_string>
  </support>
  </a>
- <a score="14.129519" run_id="dfki071ende" group_id="3601" q_id="0002">
  <answer>42</answer>
- <support>
  <s_id>SDA.950109.0187.xml</s_id>
  <s_string>Elvis Presley war 1977 im Alter von 42 Jahren gestor-
     ben.</s_string>
    </support>
    </a>
-    <a    score="0.009016844"    run_id="dfki071ende"    group_id="3602"
     q_id="0003">
  <answer>USA</answer>
- <support>
  <s_id>de_xml/p09/984837.xml</s_id>
  <s_string>Henry Martin war ein Freimaurer; zusammen mit Edouard René
     Lefèvre de Laboulaye, den Enkeln des Marquis de Lafayette und
     Frédéric Auguste Bartholdi, einem jungen Künstler aus dem Elsaß,
     war Henry Martin maßgeblich an der Finanzierung der Frei-
     heitsstatue, einem Geschenk an die USA beteiligt, deren Einweihung
     Henry Martin nicht mehr miterlebte.</s_string>
    </support>
    </a>
```

2.5 Evaluation

As far the evaluation process is concerned, no changes were made with respect to the 2006 edition. Human judges assessed the exact answer (i.e. the shortest string of words which is supposed to provide the exact amount of information to answer the question) as:

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

Most assessor-groups managed to guarantee a second judgement of all the runs.

As regards the evaluation measures, the main one was accuracy, defined as the average of SCORE(q) over all 200 questions q, where SCORE(q) is 1 in the first answer to q in the submission file is assessed as R, and 0 otherwise.

In addition most assessor groups computed the following measures:

- K1 [6]:
- Confident Weighted Score (CWS) [17] .

3 Results

As far as accuracy is concerned, scores were generally far lower this year than usual, as Figure 1 shows. Although comparison between different languages and years is not

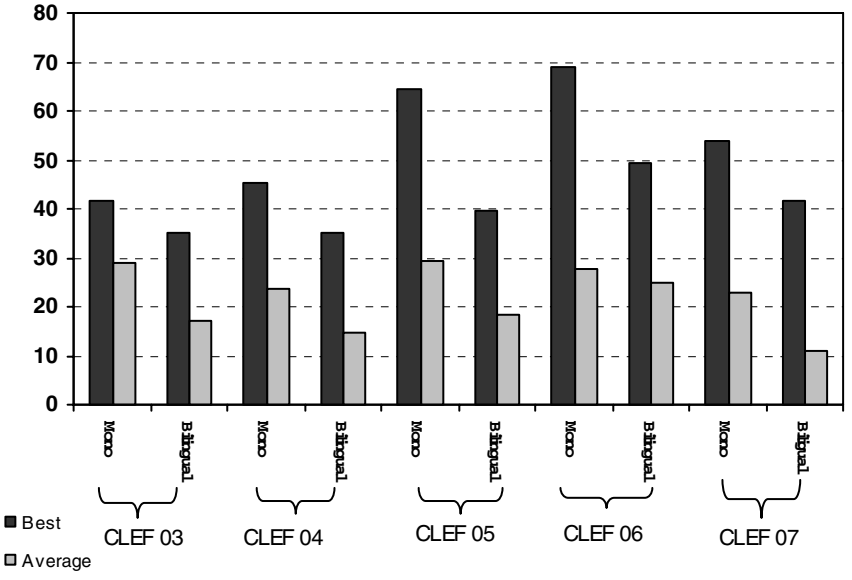


Fig. 1. Best and average scores in QA@CLEF campaigns

possible, we can observe in Figure 1 some trends this year: best accuracy both in the monolingual and the bilingual tasks decreased considerably.

This is also true for average performances. This year a neat decrease has been recorded in the bilingual tasks, due also to the presence of systems which participated for the first time, achieving very low score in tasks which are quite difficult also for veterans.

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

3.1 Participation

After years of constant growth, the number of participants has decreased in 2007 (see Table 5) due to the new challenges introduced in the exercise.

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

Table 5. Number of participants in QA@CLEF

	America	Europe	Asia	Australia	TOTAL
CLEF 2003	3	5	0	0	8
CLEF 2004	1	17	0	0	18
CLEF 2005	1	22	1	0	24
CLEF 2006	4	24	2	0	30
CLEF 2007	**3**	**17**	**1**	**1**	**22**

Table 6. Number of submitted runs

	Submitted runs	Monolingual	Cross-lingual
CLEF 2003	17	6	11
CLEF 2004	48	20	28
CLEF 2005	67	43	24
CLEF 2006	77	42	35
CLEF 2007	**37**	**23**	**14**

Also the number of submitted runs has decreased sensibly, from a total of 77 registered last year to 22 (see The geographical distribution has anyway remained almost the same, recording a new entry of a group from Australia. No participants took part to any Bulgarian tasks. Table 6). A breakdown of participants and runs, according to language, is shown in Table 4 (Section 2.3). As in previous campaigns, a larger number of people chose to participate in the monolingual tasks, which once again demonstrated to be more approachable.

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

3.2 Dutch as Target

For the Dutch subtask of the CLEF 2007 QA task, three annotators generated 200 questions organized in 78 groups so that there were 16 groups with one question, 21 groups with two, 22 with three and 19 groups with four questions. Among the 200 questions 156 were factoids, 28 definitions and 16 list questions. In total, 41 questions had temporal restrictions. Table 7 and Annotators were asked to create questions with answers either in Dutch Wikipedia or in the Dutch newspaper corpus, as well as questions without known answers. Of 200 questions, 186 had answers in Wikipedia, and 14 in the newspaper corpus. Annotators did not create NIL questions.

Table 8 below show the distributions of topic types for groups and expected answer types for questions.

Table 7. Distribution of topic types

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

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

Table 8. Distribution of expected answers for questions

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

This year, two teams took part in the QA track with Dutch as the target language: the University of Amsterdam and the University of Groningen. The latter submitted both monolingual and cross-lingual (English to Dutch) runs. The 5 submitted runs were assessed independently by 3 Dutch native speakers in such a way that each question group was assessed by at least two assessors. In case of conflicting assessments, assessors were asked to discuss the judgements and come to an agreement.

Most of the occured conflicts were due to difficulties in distinguishing between *inexact* and *correct* answers. Table 9 above shows the evaluation results for the five submitted runs (three monolingual and two cross-lingual). The table shows the number of Right, Wrong, ineXact and Unsupported answers, as well as the percentage of correctly answered Factoids, Temporally restricted questions, Definition and List questions.

The best monolingual run (gron072NLNL) achieved accuracy of 25.5%, which is slightly less that the best results in the 2006 edition of the QA task. The same tendency holds for the performance on factoid and definition questions.

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

Table 9. Results for Dutch as target

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

3.3 English as Target

160 Factoids (in groups) were requested, together with 30 definitions and ten lists. The numbers of temporally restricted factoids and questions with NIL answers was at our discretion. In the end we submitted 161 factoids, 30 definitions and nine lists. In previous years we have been obliged to devise a considerable number of temporally restricted questions and this has proved very difficult to do with the majority of them being very contrived and artificial. For this reason it was intended to set no such questions this year.

However, one reasonable one was spotted during the data entry process and so was flagged as such. Two others were also flagged accidentally during data entry. Unfortunately, therefore, the statistics cannot tell us anything about temporally restricted questions.

To achieve the goals set by the organizers it was necessary to find topics about which several questions could be asked and then to devise a set of questions from that topic. Each task was surprisingly hard, and an inevitable consequence was that the questions are much harder this year than in previous years. We had no wish to set especially difficult or convoluted questions, but unfortunately this arose as a side-effect of the new procedures.

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

Q24: *What is the "KORG" on which Niky Orellana is a soccer commentator?*

The breakdown of the questions can be summarised as follows. There were 200 questions divided into 67 groups. In other words, there were 67 initial questions (33.50%) and 133 follow-on questions (66.50%) within the collection. Reference answers were established using the three collections. Of the 236 supporting snippets included in the corpus, 88 are from the LA Times (44.00%), 68 are from the Glasgow Herald (34.00%) and 44 are from the English Wikipedia (22.00%). Thus the majority of the reference answers were in the newpapers. However, as we shall see later, some systems found a much higher proportion of answers in the Wikipedia.

Table 10. Results for English as target

Run	R #	W #	X #	#U	% F [161]	% T [3]	% D [30]	% L [9]	NIL #	% [0]	CWS	K1	Overall accuracy
cind071fren	26	171	1	2	11.18	0.00	23.33	11.11	0	0.00	0.00	0.00	13.00
cind072fren	26	170	2	2	11.18	0.00	23.33	11.11	0	0.00	0.00	0.00	13.00
csui071inen	20	175	4	1	10.56	0.00	10.00	0.00	0	0.00	0.00	0.00	10.00
dfki071deen	14	178	6	2	4.35	0.00	23.33	0.00	0	0.00	0.00	0.00	7.00
dfki071esen	5	189	4	2	1.86	0.00	6.67	0.00	0	0.00	0.00	0.00	2:50

Five cross-lingual runs with English as target were submitted this year, as compared with thirteen for last year. Five groups participated in six source languages, Dutch, French, German, Indonesian, Romanian and Spanish. DFKI submitted runs for two source languages, German and Spanish, while all other groups worked in only one. Cindi Group and Macquarie University both submitted two runs for a language pair (French-English and Dutch-English respectively) but unfortunately there was no language for which more than one group submitted a run. This means that no direct comparisons can be made between QA systems this year, because the task being solved by each was different.

An XML format was used for the submission of runs this year, by contrast with previous years when fairly similar plain text formats were adopted. This meant that our evaluation tools were no longer usable. However, last year we also participated in the evaluation of the Question Answering using Wikipedia task (WiQA)[3] organised by University of Amsterdam. For this they developed an excellent web-based tool which was subsequently adapted for this year's Dutch CLEF evaluations[4]. It allows multiple assessors to work independently, shows runs anonymised, allows all answers to a particular question to be judged at the same time (like the TREC software), and includes the supporting snippets for each submitted answer as well as the 'correct' (reference) answer. It also shows inter-assessor disagreement, and, once this has been eliminated, can produce the assessed runs in the correct XML format. Overall, this software worked perfectly for us and saved us a considerable amount of time.

All answers were double-judged[5]. Where assessors differed, the case was discussed between us and a decision taken. We measured the agreement level by two methods. For Agreement 1 we take agreement on each group of 8 answers to a question as a whole as either exactly the same for both assessors or not exactly the same. This is a very strict measure. There were disagreements for 30 questions out of the 200, i.e. 15%, which equates to an agreement level of 85%.

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

Considering all question types together, the best performance is University of Wolverhampton with 28 R and 2 X, (14% strict or 15% lenient) closely followed by the CINDI Group at Concordia University with 26 R and 1 X (13% strict or 13.50% lenient). Note that these systems are working on different tasks (RO-EN and FR-EN respectively) as noted above, so the results are not directly comparable. The best performance last year for English targets was 25.26%. Nevertheless, considering the

[3] `http://ilps.science.uva.nl/WiQA/`

[4] We are extremely grateful to Martin de Rijke and Valentin Jijkoun for allowing us to use it and for setting it up in Amsterdam especially for us.

[5] The first assessor was Richard Sutcliffe and the second was Udo Kruschwitz from University of Essex to whom we are indebted for his invaluable help.

extreme difficulty of the questions, this represents a remarkable achievement for these systems.

For Factoids alone, the best system was CINDI (FR-EN) at 11.18% followed by University of Indonesia (IN-EN) with 10.56%. For Definitions the best result was University of Wolverhampton (RO-EN) with 43.33% correct, followed equally by CINDI (FR-EN) and DFKI (DE-EN) both with 23.33%. It is interesting that this For Factoids alone, the best system was CINDI (FR-EN) at 11.18% followed by University of Indonesia (IN-EN) with 10.56%. For Definitions the best result was University of Wolverhampton (RO-EN) with 43.33% correct, followed equally by CINDI (FR-EN) and DFKI (DE-EN) both with 23.33%. It is interesting that this year the best Definition score is almost four times the best Factoid score, whereas last year they were nearly equal. One reason for this may be that the definitions either occurred first in a group of questions or on their own in a 'singleton' group. This was not specifically intended but seems to be a consequence of the relationship between Factoids and Definitions, namely that the latter are somehow epistemologically prior to the former[6]. In consequence, Definitions may be more simply phrased than Factoids and in particular may avoid co-reference in the vast majority of cases.

Nine list questions were set but only CINDI was able to answer any of them correctly (11.11% accuracy). (University of Indonesia was ineXact on one list question.) Perhaps the problem here was recognising the list question in the first place – unlike at TREC they are not explicitly flagged.

Considering the runs collectively, only 119 correct answers were returned out of 1600 attempts (8 runs and 200 questions). 70 questions were answered correctly by at least one system and thus 130 were not answered by any system. Table 11 shows a breakdown of correct answers by the collection used by a system, and by the position of an answer in a particular question group. Taking the collections first, we can see that the systems used the Wikipedia much more than we might have expected. 22% of the reference answers came from the Wikipedia (see earlier) while here we see that the total figures (excluding DFKI) are 19/100 for Glasgow Herald (19%), 14/100 for LA Times (14.%) and 67/100 for Wikipedia (67%). These figures suggest that many questions which were set relative to the newspapers were not answered from them. We will need to pay careful attention to this point before the next contest.

The last two columns in Table 11 show how good each system was at answering the first question in a group and the subsequent questions in a group – recall that there were 200 questions in 67 groups. As we can see, the number of subsequent questions answered correctly was less than the number of first questions. Across all the runs there were 68 correct answers to first questions (out of 67*8 attempts) and 51 correct answers to subsequent questions (out of 133*8 attempts). Thus the overall success rate on first questions was 12.69% and that on subsequent questions was 4.79%. This can be accounted for by the fact that subsequent questions are much more difficult because they use anaphoric references and also can involve knowing the answers to previous questions, as discussed earlier. The first column gives the number of correct answers returned by a system. The columns GH, LA and WI give the number of correct answers supported by snippets from the Glasgow Herald, LA Times and English

[6] Perhaps it is just a consequence of setting too many undergraduate examination papers!

Table 11. Breakdown of the answers by collection and by the position of the question in a group

Run	All	GH	LA	WI	First	Subsq.
cind071fren	26	4	3	19	14	12
cind072fren	26	4	3	19	14	12
csui071inen	20	6	3	11	12	8
dfki071deen	14	-	-	-	9	5
dfki071esen	5	-	-	-	3	2
mqaf071nlen	0	0	0	0	0	0
mqaf072nlen	0	0	0	0	0	0
wolv071roen	28	5	5	18	16	12

Wikipedia respectively. The last two columns show the numbers of correct answers for initial questions in a group (First) and subsequent questions in a group (Subsq.). Information about DocIDs could not be extracted or inferred from the DFKI runs.

3.4 French as Target

This year two groups took part in evaluation tasks using French as target language: one French group: Synapse Développement; and one American group: Language Computer Corporation (LCC).

In total, only two runs have been returned by the participants: one monolingual run (FR-to-FR) from Synapse Développement and one bilingual run (EN-to-FR) from LCC.

It appears that the number of participants for the French task has clearly decreased this year, certainly due to the many changes that appeared in the 2007 Guidelines for the participants: adding to a large new answer source (the static version of Wikipedia, frozen in November 2006) and adding to a large number of topic-related questions. 200 answers were assessed for syn07frfr, and 194 for lcc0707enfr.

Figure 2 shows the best scores for systems using French as target in the last four CLEF QA campaigns.

Table 12. Results for French *as target*

Run	R #	W #	X #	U #	% F [161]	% T [3]	% D [30]	% L [9]	NIL #	NIL % [0]	CWS	K1	Overall accuracy
syn07frfr	108	82	9	1	52.76	46.34	74.07	20	40	22.5	-	-	54 %
lcc07enfr	81	95	14	4	44.17	46.34	22.22	30	0	0	0.2223	-0.1235	41.75 %

The French test set was composed of 200 questions: 163 Factual (F), 27 Definition (D) and 10 closed List questions (L). Among these 200 questions, 41 were Temporally restricted questions (T).

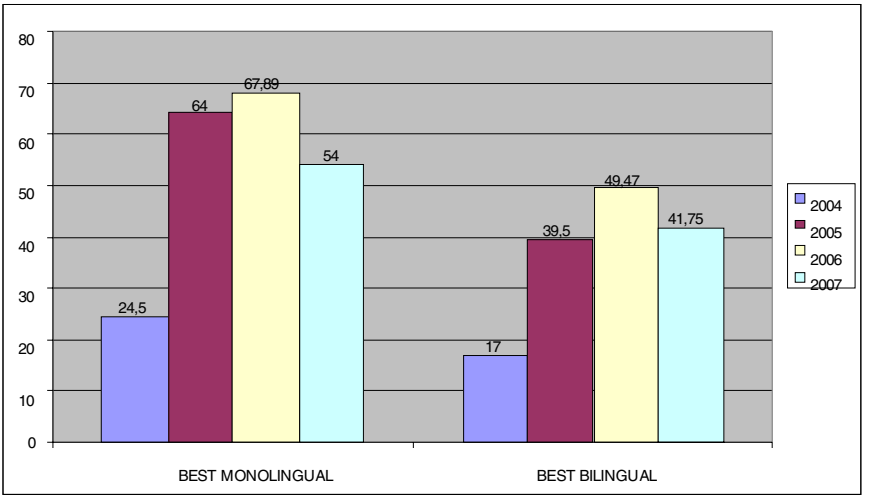


Fig. 2. Best scores for systems using French as target in QA@CLEF campaigns

The accuracy has been calculated over all the answers of F, D, T and L questions and also the Confidence Weighted Score (CWS) and the K1 measure.

For the monolingual task, the Synapse Développement' system returned 108 correct answers i.e. 54 % of correct answers (as opposed to 67,89 % last year).

For the bilingual task, the LCC's system returned 81 correct answers i.e. 41,75 % of correct answers (as opposed to 49,47 % for the best bilingual system last year).

We can observe that the two systems obtained different results according to the answer types. The monolingual system obtained better results for Definition questions (74,07 %) than for Factoid (52,76 %) and Temporally questions (46,34 %) whereas the bilingual system obtained better results for Temporally (46,34 %) and Factoid questions (44,17 %) than for Definition questions (22,22 %).

We can note that the bilingual system has not returned NIL answer, whereas the monolingual one returned 40 NIL answers (out of 9 expected NIL answers in the French test set). As there were only 9 NIL answers in the French test set and as the monolingual system returned 40 NIL answers, his final score is not very high (even if this system returned the 9 expected correct NIL answers).

In conclusion, despite the important changes in the Guidelines for the participants, the monolingual system obtained the best results of all the participants at CLEF@QA track this year (108 correct answers out of 200).

We can note that the American group (LCC) participated only for the second time in the Question Answering track using French in target and has already obtained good results that can let us imagine it will improve again in the future. In addition, we can still observe this year the increasing interest in Question Answering for the tasks using French as target language from the non-European research community due to the second participation of an American team.

3.5 German as Target

Two research groups submitted runs for evaluation in the track having German as target language: The German Research Center for Artificial Intelligence (DFKI) and the Fern Universität Hagen (FUHA).

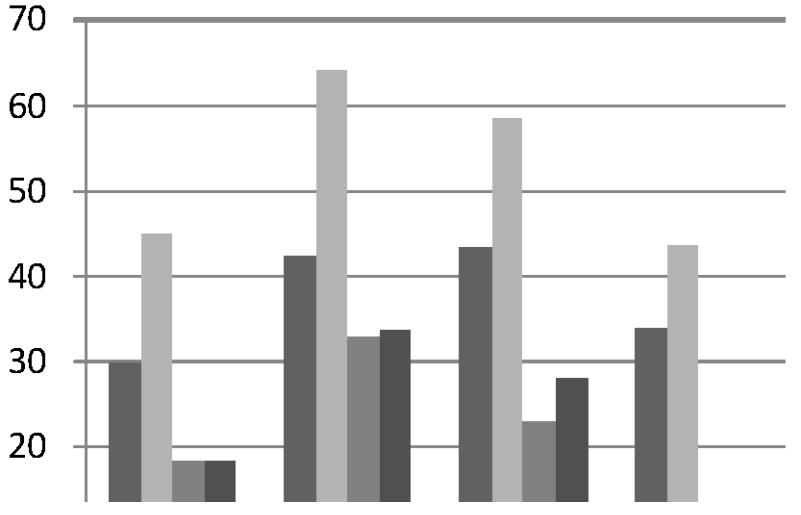


Fig. 3. Results evolution

Both provided system runs for the monolingual scenario and just DFKI submitted runs for the cross-language English-German and Portuguese-German scenario. Compared to the previous editions of the evaluation forum, this year a decrease in the accuracy of the best performing system and of an aggregated virtual system for both monolingual and cross-language tasks was registered, as seen in Figure 3.

The number of topics covered by the test set questions was of 116 distributed as it follows: 69 topics consisting of 1 question, each 19 topics of 2 and 3 related questions, and 9 topics of 4 related questions. The distribution of the topics over the document collections (CLEF vs. Wikipedia) is presented in Table 13.

Table 13. Results for German as target

Run	R #	W #	X #	U #	% F [164]	% T [27]	% D [28]	% L [8]	NIL #	NIL % [0]	CWS	K1	Overall accuracy
dfki071dede$_M$	60	121	14	5	29.8	14.81	39.29	0	0	0	-	-	30
fuha071dede$_M$	48	146	4	2	24.39	18.52	28.57	0	0	0	0.086	-0.17	24
fuha072dede$_M$	30	164	4	2	17.07	14.81	7.14	0	0	0	0.048	-0.31	15
dfki071ende$_C$	37	144	18	1	17.68	14.81	25	12,5	0	0	-	-	18.5
dfki071ptde$_C$	10	180	10	0	3.66	7.41	14.29	0	0	0	-	-	5

According to Tables 14 and 15 the most frequent topic types were PERSON (40), OBJECT (33) and ORGANIZATION (23), with first two types more present for the news collection of documents (CLEF).

Table 14. Topic distribution over data collections

Topic Size	# Topics / CLEF	# Topics / WIKI	# Topics
1	53	16	69
2	4	15	19
3	4	15	19
4	7	2	9
Total	68	48	116

As regards the source of the answers, 101 questions from 68 topics asked for information out of the CLEF document collection and the rest of 99 from 48 topics for information from Wikipedia. Table 16 shows a breakdown of the test set questions by the expected answer type (EAType) for each collection of data.

Table 15. Topic type breakdown over CLEF collection

Topic Type	Topic Size				Total
	1	2	3	4	
PERSON	23	2	0	3	28
OBJECT	19	0	1	0	20
ORGANIZATION	8	1	1	2	12
LOCATION	1	1	1	0	3
EVENT	2	0	1	2	5
OTHER	0	0	0	0	0
					68

The system developed by DFKI relies on shallow NLP methods for both question and document processing and uses distance-based metrics and recall evidence for answer selection. The system developed by FUHA combines both shallow and deep NLP methods and uses semantic representations and an entailment engine for answer selection.

The details of systems' results can be seen in Table 13. There were no NIL questions tested in this year's evaluation. The results submitted by DFKI did not provide a normalized value for the confidence score of an answer and therefore both CWS and K1 values could not be computed.

A breakdown of results along self-contained questions, i.e. first ones in a topic with no reference to previous stated information – 116 in total, and linked questions, i.e.

Table 16. Topic type breakdown over Wikipedia collection

Topic Type	Topic Size				Total
	1	2	3	4	
PERSON	4	2	5	1	12
OBJECT	5	5	3	0	13
ORGANIZATION	3	3	5	0	11
LOCATION	2	1	1	1	5
EVENT	2	3	1	0	6
OTHER	0	1	0	0	1
					48

questions related to previous mentioned information or to the topic – 84 in total, shows a drop in the systems' accuracy for the latter.

A thorough analysis of the questions unanswered by any of the participating systems revealed following common features of them:

- The answer's context covers at least two sentences that might be adjacent (CLEF collection) or not (Wikipedia collection).
- The question and the answer's context share semantic items, i.e. concepts, but not lexical items, i.e. words. Some examples of this phenomena are:
  - Ehe (marriage) vs verheiratet (married)
  - Geburtsname (birth name) vs bürgerlicher Name (civil name)
  - Band vs Popgruppe
  - Spielfilm von (motion picture by) vs verfilmt von (filmed by)
  - Beruf (profession) vs Rechtsanwalt (lawyer)
- The date asked for in question is not explicitly mentioned in the answer's context, but assumed based on document's publication date.

The assessment was conducted by two native German speakers with fair knowledge of information access systems. Table 17 describes the inter-rater disagreement on the assessment of answers in terms of question and answer disagreement. Question disagreement reflects the number of questions on which the assessors delivered different judgments. Along the total figures for the disagreement, a breakdown at the

Table 17. Inter-assessor agreement/disagreement (breakdown)

Run	Number of questions	# Q-Disagreements						
		Total	F	D	L	X	U	W/R
dfki071dede$_M$	200	20	16	4	0	15	4	1
fuha071dede$_M$	200	13	10	3	0	7	3	3
fuha072dede$_M$	200	7	6	1	0	2	2	3
dfki071ende$_C$	200	13	7	5	1	12	1	0
dfki071ptde$_C$	200	8	3	5	0	8	0	0

question type level (Factoid, Definition, List) and at the assessment value level (ineXact, Unsupported, Wrong/Right) is listed. The answer disagreements of type Wrong/Right are trivial errors during the assessment process when a right answers was considered wrong by mistake and the other way around, while those of type X or U reflect different judgments whereby an assessor considered an answer inexact or unsupported while the other marked it as right or wrong.

3.6 Italian as Target

Only one group took part this year in the monolingual Italian task, i.e. FBK-irst, submitting only one run. The results are shown in Table 18.

Table 18. Results of the Italian monolingual task

Run	R #	W #	X #	U #	% F [161]	% T [3]	% D [30]	% L [9]	NIL Returned	NIL Correct	CWS	K1	Overall accuracy
irst071itit	23	160	4	13	15.17	12.5	2.63	0	14	3	0.017	0.043	11.55%

The Italian question set consisted of 147 factoid questions, 41 definition questions and 12 list questions. 38 questions contained a temporal restriction, and 11 had no answer in the Gold Standard. In the Gold Standard, 108 answers were retrieved from Wikipedia, the remains from the news collections (see Table 21). Results for Italian as target (answers to linked and unlinked questions). As Table 19 shows, the question set was almost perfectly balanced between questions were linked to a topic –which could contain co-references and needed to considered as a group- and self-contained questions –which were similar to the queries proposed in the previous campaigns.

The submitted run was assessed by two judges; the inter-annotator agreement was 92,5%, meaning that the dataset contained a very low percentage of questionable cases.

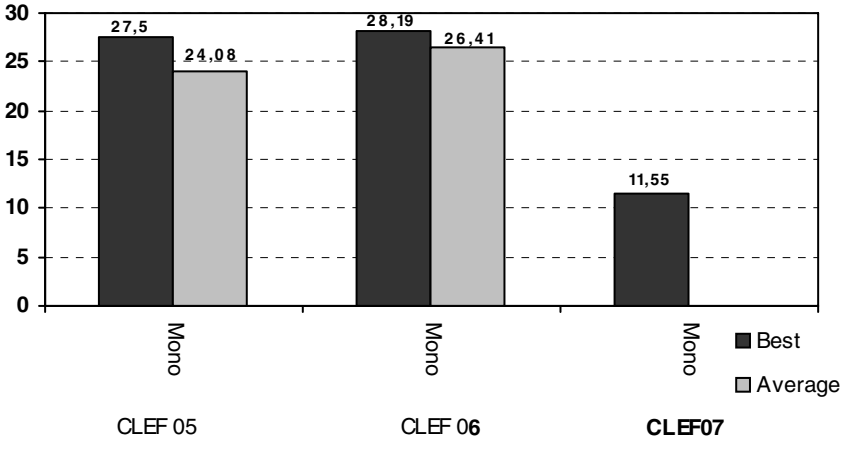


Fig. 4. Best and Average performance in the Monolingual and Bilingual tasks

As Figure 4 shows, the performance of the FBK-irst system was lower than those achieved in the previous campaigns: in 2006 the accuracy in the monolingual task was 22.87, almost twice as much as this year's score. Anyway, these results reflected the general trend also of the performances of the other systems largely due to the innovations introduced.

The system achieved low accuracy in all types of questions, performing somehow better in factoids questions. Definition questions, with 2.63% of accuracy and list questions, for which no correct answer was retrieved, proved to be particularly challenging.

Table 19. Results for Italian as target (answers to linked and unlinked questions)

	#	%	R	W	X	U
Question linked to a topic	108	54%	0	106	0	2
Self-contained questions	92	46 %	23	54	4	11
Total	200	100%	23	160	4	13

A relevant number of questions (about 6%) was judged unsupported, meaning that the correct answer was retrieved by the system, which did not provided enough context to support it.

Table 20. Results for Italian as target for NIL questions

	Precision (Overall)	**Recall (Overall)**
FBK-irst	0.21	0.27

Regarding the questions with no answers, the system returned the value NIL 14 times, compared to the 11 present in the Gold Standard. Therefore, as Table 20 shows, the overall precision about NIL questions was 0.21, with an overall recall of 0.27, which proves that NIL questions are still problematic.

It may be interesting to have a closer look at the results according to the new features introduced in this year's competition.

Table 21. Questions by source

Source	**# Gold Standard**	**# FBK-irst**	**FBK-irst %**	**R**	**W**	**X**	**U**
News	81	178	89%	19	142	4	13
Wikipedia	108	8	4%	1	7	0	0
Other (NIL)	11	14	7%	3	11	0	0
Total	200	200	100%	23	160	4	13

Meanwhile the answers in the Gold standard were almost equally retrieved from news collections and Wikipedia (see Table 21), the system found the answers mainly in the news collections, for a total of 178 out of 200, compared to the 8 responses extracted from Wikipedia. If we consider that in the Gold standard the answer retrieved from Wikipedia were 108, we could conclude that the system did not exploit this source properly. The reason for that should be probably investigated a bit longer. As for the precision of the answer with respect to the collections, it was 0.13 on Wikipedia and 0.11 on the news collection.

3.7 Portuguese as Target

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

Again, Priberam presented the best result for the third year in a row; the group of the University of Évora wasn't however far behind. As last year, we added the classification X-, meaning incomplete, while keeping the classification X+ for answers with extra text or other kinds of inexactness. In Table 22 we present the overall results.

A direct comparison with last year's results is not fully possible, due to the existence of multiple questions to each topic. Therefore, in Question 94 was reclassified as NIL due to a spelling error, and question 135 because of the use of a rare meaning of a word. On the other hand, one system saw through that rare meaning, providing a correct answer; we decided to keep the question as NIL, considering correct both the system's answer and any NIL answer from other systems.

Table 23 we present the results both for first question of each topic (which we believe is more readily comparable to the results of previous years) and for the linked questions.

As it can be seen, apart from Priberam, the results over linked questions aren't much different from those over not-linked. On the whole, compared to last year [12], Priberam saw a slight drop on its results, Raposa (FEUP) a clear improvement from an admitedly low level, Esfinge (SINTEF) a clear drop, and LCC kept last year's levels. Senso (UE) shows a marked improvement since its last participation in 2005 [16].

Table 22. Results for Portuguese as target (all 200 questions)

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

Question 94 was reclassified as NIL due to a spelling error, and question 135 because of the use of a rare meaning of a word. On the other hand, one system saw through that rare meaning, providing a correct answer; we decided to keep the question as NIL, considering correct both the system's answer and any NIL answer from other systems.

Table 23. Results for Portuguese as target (answers to linked and unlinked questions)

Run	First questions [149]						Linked questions [51]	
	R	W	X+	X-	U	Accuracy	R	Accuracy
	#	#	#	#	#	%	#	%
diue071ptpt	61	77	1	9	1	40.9	23	45.1
esfi071ptpt	11	132	0	4	2	7.4	5	9.8
esfi072ptpt	6	141	0	1	1	4.0	6	11.8
feup071ptpt	34	113	1	1	0	22.8	6	11.8
ines071ptpt	17	125	1	4	2	11.4	5	9.8
ines072ptpt	21	122	0	4	2	14.1	7	13.7
prib071ptpt	92	86	3	5	1	61.7	9	17.6
lcc_071enpt	44	48	7	3	9	29.5	12	23.5

The same system also found a correct answer to a question classified as NIL in the test set; that question was therefore reclassified as non-NIL. In the end, there were 13 NIL questions. Table 24 shows the results for each answer type of definition questions, while Table 25 shows the results for each answer type of factoid questions (including list questions). As it can be seen, four out of six systems perform clearly better when it comes to definitions than to factoids. This may well have been helped by the use of Wikipedia texts, where a large proportion of articles begin with a definition.

We included in both Table 24 and Table 25 a virtual run, called combination, in which one question is considered correct if at least one participating system found a valid answer. The objective of this combination run is to show the potential achievement when combining the capacities of all the participants. The combination run can be considered, somehow, state-of-the-art in monolingual Portuguese question answering. The system with best results, Priberam, answered correctly 72.7% the questions with at least one correct answer, not as dominating as last year; in all, 137 questions had at least one correct answer among the monolingual runs (67% of first questions and 47% of linked questions); 75 questions were answered by more than one system, and only four (all NIL) by all monolingual runs.

Despite being a bilingual run, LCC answered correctly to 14 questions not answered by any of the monolingual systems.

Analysing those questions which no system managed to answer, and comparing them with the test set extract chosen by the organization, it seems that the most important cause are the non-handling of anaphora - both in the questions (while only 20% of the first questions of each topic found no correct answer, that number rises to 37% of the subsequent questions), and of the collection text itself (e.g., questions 170 and 172).

Table 24. Results of the assessment of the monolingual Portuguese runs: definitions

Run	obj	org	oth	per	TOT	%
	5	22	28	24		
diue071ptpt	6	4	5	4	19	63%
esfi071ptpt	1	0	0	0	1	3%
esfi072ptpt	1	0	0	0	1	3%
feup071ptpt	3	2	4	7	16	53%
ines071ptpt	4	4	6	0	14	47%
ines072ptpt	5	5	6	2	18	60%
prib071ptpt	6	4	6	7	23	77%
combination	6	5	8	9	27	87%
lcc_071enpt	2	3	2	1	8	27%

Secondary issues yet to be fully tackled are the use of the document date when not mentioned in the document (e.g., questions 71, 128, 168 and 194), of using the common root of words to find an answer (questions 11 and 196), of validating dates when intervals are used in the text (questions 55 and 140) and of finding Portuguese equivalents for Brazilian terms or vice-versa (questions 21 and 98).

Table 25. Results of the assessment of the monolingual Portuguese runs: factoids, including lists

Run	cou	loc	Mea	obj	org	oth	per	tim	TOT	%
	21	34	16	5	22	28	24	20		
diue071ptpt	11	17	4	3	6	8	7	9	65	38%
esfi071ptpt	3	3	0	0	1	0	1	7	15	9%
esfi072ptpt	2	4	0	0	1	0	2	2	11	6%
feup071ptpt	4	8	0	0	3	1	3	5	24	14%
ines071ptpt	1	3	0	0	0	0	2	2	8	5%
ines072ptpt	2	4	0	0	0	0	2	2	10	6%
prib071ptpt	9	15	10	1	11	14	8	10	78	46%
combination	16	24	12	3	12	17	12	13	109	64%
lcc_071enpt	7	11	6	1	3	10	4	6	48	28%

In Table 27 presents the results of the 20 temporally restricted questions. As in previous years, the effectiveness of the systems to answer those questions is visibly lower than for non-TRQ questions (and indeed most systems only answered correctly question 160, which is a NIL TRQ).

Table 26 we present some values concerning answer and snippet size. Table 27 presents the results of the 20 temporally restricted questions. As in previous years, the effectiveness of the systems to answer those questions is visibly lower than for non-TRQ questions (and indeed most systems only answered correctly question 160, which is a NIL TRQ).

Table 26. Average size of answers (values in number of words)

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

A total of twelve questions were defined as list questions; unlike last year, all these questions were closed list factoids, with two to twelve answers each[7].

The results were, in general, weak, with UE and LCC getting two correct answers, Priberam five, and all other system zero. There was a single case of incomplete answer (i.e., answering some elements of the list only), but it was judged W since, besides incomplete, it was also unsupported.

Table 27. Accuracy of temporally restricted questions

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

Table 28 presents the distribution of questions by source during their selection, while Table 29 presents the distribution of sources used by the different runs and their correctness.

As it can be seen, the systems found the answers to half of the questions originally selected from newswire in Wikipedia (27 out of 55); conversely, only 5% of questions selected from Wikipedia received a correct answer from newspaper sources.

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

Table 28. Questions by source

Source	# during selection	# including valid answers
Wikipedia	132	159
News	55	62
NIL	13	13

Table 29. Answers by source and their correctness

Run	News #	% correct	Wikipedia #	% Correct	NIL #	% correct
diue071	10	80%	79	81%	111	11%
esfi071	53	9%	4	50%	143	6%
esfi072	18	6%	1	0%	181	6%
feup071	17	71%	39	44%	144	8%
ines071	30	13%	23	39%	147	6%
ines072	28	14%	23	57%	149	7%
prib071	41	63%	141	50%	18	28%
lcc_071	17	24%	174	30%	9	0%

3.8 Romanian as Target

The creation of the questions was realized at the Faculty of Computer Science, Al.I. Cuza University of Iasi. The group[8] was very well instructed with respect to this task, using the Guidelines for Question Generation and based on a good feedback received from the organizers at IRST[9]. The final 200 created questions are distributed according to Table 30 where for each type of question and expected answer we indicate also the temporally restricted questions out of the total number of questions. For Romanian, as source and target language we used only the collection of Wikipedia articles, hence the answers of 100% of the questions are in Wikipedia (without counting the NIL questions).

Table 30. Question and answer types distribution in Romanian (in brackets the number of temporally restricted questions)

Q type / expected A type	PERS	TIME	LOCAT	ORG	MEA	COU	OBJ	OTH	TOTAL
FACTOID	22 (14)	17	21 (4)	19 (8)	17	20 (7)	16 (6)	21 (6)	153 (45)
DEFINITION	9	-	-	5	-	-	6 (1)	10 (1)	30 (2)
LIST	5 (1)	-	2	-	-	-	1	2	10 (1)
NIL	3 (1)	-	-	-	-	1 (1)	2 (1)	1	7 (3)

[8] Three Computational Linguistics Master students: Anca Onofraşc, Ana-Maria Rusu, Cristina Despa, supervised and working in collaboration with the two organizers.

[9] Without the help received from Danilo Giampicolo and Pamela Forner, we wouldn't have solved all our problems.

We decided to include NIL questions, even though they seem somehow unnatural; the way we created them was not by including questions about facts impossible from a human perception. The Romanian NIL questions have answers in the English online Wikipedia, but not in the frozen Romanian Wikipedia articles.

This year in QA@CLEF one novelty were the questions related under the same topic: the organizers had to choose a certain number of topics and to create up to four questions related under one same topic. Using also the classification available within the question generation upload interface[10], the percentage of topic-linked questions is illustrated in Table 31. This table shows that 129 questions were grouped under 51 topics, hence 64.5% out of the total 200 questions were linked in under topics with more than one question.

Most difficulties in this task were raised by deciding on the supporting snippets, especially for questions belonging to the same topic. We found unnatural to include answers through "copy-paste" from the text, because this way the answer was grammatically incorrect in some situations.

Table 31. Percentage of topic-linked questions

# of questions / Topic type	PERSON	LOCATION	ORGANIZ.	EVENT	OBJECT	OTHER	TOTAL
4 Qs	4		1			1	6
3 Qs	6	1	1		4	3	15
2 Qs	11	5	4	2	3	5	30
1 Q	14	7	15	3	11	21	71
TOTAL	**35**	**13**	**21**	**5**	**18**	**30**	**122**

For the LIST question we prepared also some questions with the answer to be found in various sections of an article or even in various articles. The situation is plausible from the point of view of an user asking for automatic answers.

We illustrate only the first type of LIST question with the following example: for the question *Name the main laws initiated by Cuza.* (RO: *Numiți prinipalele legi inițiate de Cuza.*), the answer should be extracted from various sentences in the same article[11]. We show (underlined) only the sentences from where the answer should be extracted: *[…] se întocmeşte un Proiect de lege organică pentru instrucţia publică în Principatele Unite, […] Noul guvern prezintă Adunării şi realizează proiectul legii privind secularizarea averilor mânăstireşti, lege prin care s-a dat o lovitură puternică feudalismului. De asemenea, se supune poporului, spre aprobare prin plebiscit, o nouă contribuţie, o nouă lege electorală. […] În acest an se decretează Legea Rurală, prin care se desfiinţează iobăgia. Reforma agrară din 1864, a cărei aplicare s-a încheiat în linii mari în 1865, a satisfăcut în parte setea de pământ a ţăranilor, […].* The English version[12] of the same Wikipedia article includes even more laws: *His first measure*

[10] http://www.celct.it/Question_generation_interface/question_generation_interface.html

[11] /ro/a/l/e/Alexandru_Ioan_Cuza_9c42.html

[12] http://en.wikipedia.org/wiki/Alexandru_Ioan_Cuza

addressed a need for increasing the land resources and revenues available to the state, by "secularizing" (confiscating) monastic assets (1863). […] The land reform, liberating peasants from the last corvées, freeing their movements and redistributing some land (1864), was less successful. […] His plan to establish universal manhood suffrage, together with the power of the Domnitor to rule by decree, passed by a vote of 682,621 to 1,307. He consequently governed the country under the provisions of Statutul dezvoltător al Convenției de la Paris ("Statute expanding the Paris Convention"), an organic law adopted on July 15, 1864. With his new plenary powers, Cuza then promulgated the Agrarian Law of 1863. […]Cuza's reforms also included the adoption of the Criminal Code and the Civil Code based on the Napoleonic code (1864), a Law on Education, establishing tuition-free, compulsory public education for primary schools. The examples show that the Romanian version includes 5 answers whereas the English one has 9 laws to be included in a list answer.

This year two Romanian groups took part in the monolingual task with Romanian as a target language: the Faculty of Computer Science from the Al. I. Cuza University of Iasi (UAIC), and the Research Institute for Artificial Intelligence from the Romanian Academy (RACAI), Bucharest. Three runs were submitted – one by the first group and two by the second group [14], with the differences between them due to the way they treated the question-processing and the answer-extraction.

The RACAI systems are based on the parse tree of the candidate sentence and are using different heuristics to match keywords from the questions with those of the sentence; they use the same corpus processing tool, TTL [7] - for tokenization, POS-tagging, lemmatization, NE recognition and chunking, LexPar [8] - for link analysis, the same text search engine (based on Lucene[13]) and different question analysis and answer extraction modules.

The UAIC system follows the traditional QA systems architecture: a corpus pre-processing module, a question analyser (including an anaphora resolution (AR) module, to handle topic-related questions), a module dedicated to index creation and Information Retrieval (based on the same Lucene), and an answer extractor. Next to the AR module, another novelty of the UAIC system is the use of a Textual Entailment module for the answer extraction.

The 2007 general results are presented in Tables 32, 33 and 34. The statistics includes a system, named *combined (0)*, obtained through the combination of the 3 participating RO-RO systems. This "ideal" system permits to calculate the percentage of the questions (and their type), answered by at least one of the three systems.

Table 32. Results in the monolingual task. Romanian as target language (I).

Run	R	W	X	U	Overall accuracy	NIL returned	NIL correct
combined (0)	81	91	37	1	40.5	7	7
outputRoRo (1)	24	171	4	1	12	100	5
ICIA071RORO (2)	60	105	34	1	30	54	7
ICIA072RORO (3)	60	101	39	0	30	54	7

[13] http://lucene.apache.org/

All three systems crashed on the LIST questions. The two RACAI systems did not include rules to handle this type of question [14], whereas the UAIC system had a simple rule (if the question focus is a plural noun, then the question type is LIST).

Table 33. Results in the monolingual task. Romanian as target language (II).

Run	Factoid Questions					List Questions					Definition Questions				
	R	W	U	X	ACC	R	W	U	X	ACC	R	W	U	X	ACC
(0)	*52*	*76*	*1*	*84*	*33.98*	*0*	*10*	*0*	*0*	*0*	*22*	*5*	*0*	*3*	*73.33*
(1)	24	131	1	2	15	0	10	0	0	0	0	30	0	0	0
(2)	38	90	1	31	23.75	0	10	0	0	0	22	5	0	3	73.33
(3)	38	86	0	36	23.75	0	10	0	0	0	22	5	0	3	73.33

The NIL questions are hard to classify, starting from the question-classifier (the classifier should "know" that the QA system has no possibility, no knowledge to find the answer). It would be better to have a clear separation between the NIL answers due to impossibility to find answer and the NIL answers classified as such by the system. The performance of all the three systems with respect to the NIL questions is as high as indicated in Table 34 because the systems treated the questions non-classifiable in any of the other types (F, D or L) as NIL.

Table 34. Results in the monolingual task. Romanian as target language (III).

Run	Temporally Restricted					NIL				
	R	W	U	X	ACC	R	W	U	X	ACC
(0)	*19*	*24*	*0*	*8*	*37.25*	*7*	*0*	*0*	*0*	*100*
(1)	11	39	0	1	21.57	5	95	0	0	5
(2)	10	31	0	10	19.61	7	47	0	0	12.96
(3)	10	31	0	10	19.61	7	47	0	0	12.96

For the DEFINITION questions the UAIC system considered them as such if the expected answer is of type D, whereas the answer classifier is based on patterns, specific for each type of answer. The RACAI systems are using dedicated rules for the D questions, hence the performance is understandable. The D answers judged as X or W are due to too long answers, too short snippets or to snippets that are shortened as such as they do not include the Right answer. For example for the question *Ce este Selena?* (EN: *What is Selene?*), the answer returned by the RACAI systems was: *o actriță şi cântăreață americană , născută pe 24 iulie 1969 , în cartierul Bronx din New York* (EN: *an American actress and singer, born on July 24, 1969 in Bronx, New York*). The answer is considered "good enough" [14], but it was judged as wrong because it indicates the actress who played the role of Selena in the homonymous movie. The correct answer is *satelitul natural al Pamântului* (EN: *the natural satellite of the Earth*). The answer returned by the systems could reply to the D question "*What is Jennifer López?*", according to the sentence in the wikipedia article and the

provided snippet *Jennifer López este o actriță și cântăreață americană, născută pe 24 iulie 1969, în cartierul Bronx din New York* (EN: *Jennifer López is an American actress...*). The focus of the question (also the topic of the group of questions) is *Selena*, which anyway is not a defined entity in the text *Dar succesul a fost de partea ei abia în anul 1997, când a jucat rolul binecunoscutei și regretatei Selena, în filmul cu același nume.* (EN: *But her success came only in 1997 when she played the role of Selena, the famous and regretted person in the homonymous movie*). The topic of *Selena* proved to be for the RACAI systems a very good example of 3 topic-related questions for which the systems returned Right answers for the 2nd and 3rd questions, even though the first one had a wrong answer. The same situation appeared in many other topic-related questions answered by the RACAI systems, as we will show below. This proves that the strategy employed[14] (adding to the query generated for a new question the query of the first question of the group, namely the topic of the question and the focus of the 10 first answers returned to the previous question of the group) is a good one.

The topic-related questions were handled by UAIC through a dedicated AR module able to work by identifying the antecedents of anaphors that refer to a previous question answer or focus or by expanding the keywords lists of the questions in a same group with the keywords of the first question in that group. This strategy allowed identifying an answer as X in one case, as U in another one and as R in 9 cases of topic-related questions (the first one in the group is excluded). In 5 of these cases the R answer is NIL, hence the AR strategy was not used. For the other 4 cases the answer was R because the strategy worked and the system has specially developed rules for the MEASUREMENT answers (one case), for the temporally restricted questions (one case) or the question contains many keywords. Therefore the UAIC percentage of R answers for linked questions is 6.97%. The RACAI strategy for linked questions conducted to 20 R answers (15.5%), 15 of type X (11.62%) and 1 – U for the 2nd, 3rd or even the 4th question in a topic-related group. Six of the 20 R answers are for NIL questions, hence no strategy was used but only for the other 14 questions. One very nice such example has the topic *International Monetary Fund*, where the first question (*Which organization was formed in 1945 with the purpose of promoting a healthy global economy?*) included the topic only in the expected answer, not found by the RACAI systems. But for the second question (*How many members does it have?*) the answer is right (*184*).

The RACAI answers were judged as X or U, and not only for the topic-related questions, mainly due to answers that are too long, snippets shortened as such as they do not contain the answer (in fact in some situations the answer is the only one missing from the snippet) or because there are cases where the answer and the snippet has no connections (the answer extraction module). The UAIC answers of type X and/or U were judged as such mainly because the snippets are too long and they do not contain full clauses, but segments of clauses or sentences, unexpectedly stopped.

Due to time restrictions, all three runs where judged by only one assessor at the Faculty of Computer Science in Iasi, so an inter-annotator agreement was not possible. Based on the Guidelines, all three systems were judged in parallel. The same

[14] We thank to prof. Dan Tufis for clarifying the methodology.

evaluation criteria, especially with respect to the U and X answers, were used. The analyses described above are based on a thorough manual introspection.

3.9 Spanish as Target

The participation at the Spanish as Target subtask has decreased from 9 groups in 2006 to 5 groups this year. All the runs were monolingual. We think that the changes in the task (linked questions and Wikipedia) led to a lower participation and worse overall results because systems could not be tuned on time.

Table 35 shows the summary of systems results with the number of Right (R), Wrong (W), Inexact (X) and Unsupported (U) answers. The table shows also the accuracy (in percentage) of factoids (F), factoids with temporal restriction (T), definitions (D) and list questions (L). Best values are marked in bold face.

All the runs were assessed by two assessors. Only a 1.5% of the judgements were different and the resulting kappa value was 0,966, which corresponding to "almost perfect" assessment [10].

Table 35. Results for Spanish as target

Run	R #	W #	X #	U #	% F [115]	% T [43]	% D [32]	% L [10]	NIL #	NIL F [8]	CWS	K1	Overall accuracy
Priberam	89	87	3	21	47,82	23,25	68,75	20	3	0,29	-	-	44,5
Inaoe	69	118	7	6	28,69	18,60	87,50	-	3	0,12	0,175	-0,287	34,5
Miracle	30	158	4	8	20	13,95	3,12	-	1	0,07	0,022	-0,452	15
UPV	23	166	5	6	13,08	9,30	12,5	-	1	0,03	0,015	-0,224	11,5
TALP	14	183	1	2	6,08	2,32	18,65	-	3	0,07	0,007	-0.34	7

Table Table 36 shows some evidence on the effect of Wikipedia in the performance. When the answer appears only in Wikipedia the accuracy is reduced in more than 35% in all the cases. Regarding NIL questions, The correlation coefficient r between the self-score and the correctness of the answers (shown in Table 39), has been similar to the obtained last year, being not good enough yet, and explaining the low results in CWS and K1 [6] measures.

Table 37 shows the harmonic mean (F) of precision and recall for self-contained, linked and all questions.

The best performing system has decreased their overall performance with respect to the last edition (see Table 38). in NIL questions. However, the performance considering only self-contained questions is closer to the one obtained last year.

The correlation coefficient r between the self-score and the correctness of the answers (shown in Table 39), has been similar to the obtained last year, being not good enough yet, and explaining the low results in CWS and K1 [6] measures.

Table 36. Results for self-contained and linked questions, compared with overall accuracy

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

Table 37. Results for Spanish as target for NIL questions

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

Since a supporting snippet is requested in order to assess the correctness of the answer, we have evaluated the systems capability to extract the answer when the snippet contains it.

Table 38. Evolution of best results for NIL questions

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

The first column of Table 39 shows the percentage of cases where the correct answer was present in the snippet and correctly extracted. This information is very useful to diagnose if the lack of performance is due to the passage retrieval or to the answer extraction process. As shown in the table, the best systems are also better in the task of answer extraction, whereas the rest of systems still have a lot of room for improvement.

Table 39. Answer extraction and correlation coefficient (r) for Spanish as target

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

4 Final Analysis

This year the task was changed considerably and this affected the general level of results and also the level of participation in the task. The grouped questions could be regarded as more realistic and more searching but in consequence they were much more difficult. The policy of not declaring the question type means that if this is deduced incorrectly then the answer is bound to be wrong. Moreover, the policy of not even declaring the topic of a question group, but leaving it implicit (usually within the first question) means that if a system infers the topic wrongly, then all questions in the group will be answered wrongly. Neither of these strike us as particularly 'realistic'. In a real dialogue, if a question is answered inappropriately we do not dismiss all subsequent answers from that person, we simply re-phrase the question instead. The level of ambiguity concerning question type in a real dialogue is not fixed at some arbitrary value but varies according to many factors which the questioner estimates. In CLEF we are not modelling this process at all accurately and this affects the validity of our results. In addition, co-reference has now entered CLEF. This is interesting and useful but it might be preferable if we could separate the effect of co-reference resolution from other factors in analysing results. This could be done by marking up the co-references in the question corpus and allowing participants to use this information under certain circumstances. Finally, we have for the first time used the Wikipedia as a source of questions. For English targets there were few questions intended to be answered from it, but in practice many of the returned answers were supported by Wikipedia snippets. We could interpret this in different ways. On the one hand, we could argue that it shows how good Wikipedia is at answering simple questions, from which it follows that the newspaper corpora could be discarded. An alternative point of view, however, could be that it is valuable to be able to extract *additional* knowledge from newspapers and that therefore the Wikipedia could be excluded from certain tasks. This is a point which needs further discussion.

From the analyses accomplished by the organizing groups for German, Portuguese and Spanish, an overall decrease in the accuracy reached by the systems when treating linked questions can be observed. This fact evidences that topic resolution seems to be a weak point for QA systems. In the present edition topic-related questions were proposed for the first time and the participants did not have much time to tune their systems. As a consequence, they could not manage as well as in previous editions. There exist evidences that the most important cause is the non-handling of anaphora,

as referred the team in charge of Portuguese after an analysis of the data related to their language. From the questions which no system managed to answer for Portuguese as target language, only 20% of the first questions of each topic found no correct answer. But, that number rises to 37% of the subsequent questions.

Another source of difficulties, as referred by some participants, is the inclusion of Wikipedia as document corpus. These participants argue that the overall decrease in the accuracy reached by their systems comes from several problems when consulting Wikipedia. In all cases, these problems are a consequence of the impossibility of tuning the systems to the new requirements of the task in the time available. As instance, Synapse [11] could not adapt its system to a pattern extraction from Wikipedia as accurate as the one implemented for the news corpus. University of Hagen [5] found problems when treating article names, which led to an inconsistent concept index that rendered many Wikipedia articles inaccessible for its system.

In addition, the drop of the number of participating teams caused that, for certain pairs of source and target languages, one team tackled the subtask. Therefore, a comparison between systems working under the same circumstances cannot be accomplished. It impedes one of the major goals of campaigns such the QA@CLEF: the systems comparison in order to determine better approaches.

5 Future Work

After 5 years experiencing with QA issues, a lot of resources and know-how is accumulated nowadays. But systems do not show a brilliant overall performance, even those that participate edition by edition. The systems evidenced that they could not manage suitably the challenges proposed in the present edition while improving their performance when tackling issues already treated in previous campaigns. Given this situation, perhaps is time for no more innovation in question and answer types but for revising, little by little, every aspect considered until now in the past campaigns, in order to stimulate the improvement of the systems in a few skills every year. For this, without forgetting that nowadays sufficient evaluation resources from the previous years are available, in following campaigns a new focusing could be given to the task, as instance:

- Component evaluation, i.e., question classification, topic resolution, passage retrieval, answer extraction or answer validation (the latter already developed in the AVE).
- Join some target languages into a single multilingual target collection. Portuguese and Spanish are good candidates since they are closed languages and have many participants.
- Evaluation of an only question type every year.

In addition, being the development of high-performance QA systems a desirable goal, not only an accurate definition of every task should be accomplished but a more in-depth analysis of the participant systems, in order to determine relations between implementations and results.

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

Jesús Herrera has been partially supported by the by the Spanish Ministry of Education and Science (TIN2006-14433-C02-01 project).

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

Paulo Rocha was supported by the Linguateca project, jointly funded by the Portuguese Government and the European Union (FEDER and FSE), under contract ref. POSC/339/1.3/C/NAC.

References

1. QA@CLEF Website, http://clef-qa.itc.it/
2. AVE Website, http://nlp.uned.es/QA/ave/
3. QAST Website, http://www.lsi.upc.edu/~qast/
4. QA@CLEF 2007 Organizing Committee. Guidelines (2007), http://clefqa.itc.it/2007/download/QA@CLEF07_Guidelines-for-Participants.pdf
5. Hartrumpf, S., Glöckner, I., Leveling, J.: University of Hagen at QA@CLEF 2007: Coreference Resolution for Questions and Answer Merging. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2008)
6. Herrera, J., Peñas, A., Verdejo, F.: Question Answering Pilot Task at CLEF 2004. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 581–590. Springer, Heidelberg (2005)
7. Ion, R.: Word Sense Disambiguation Methods Applied to English and Romanian. PhD thesis, Romanian Academy, Bucharest (2007)
8. Ion, R., Mititelu, V.B.: Constrained Lexical Attraction Models. In: Nineteenth International Florida Artificial Intelligence Research Society Conference, pp. 297–302. AAAI Press, Menlo Park (2006)
9. Jijkoun, V., de Rijke, M.: Overview of the WiQA Task at CLEF 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 265–274. Springer, Heidelberg (2007)
10. Landis, J.R., Koch, G.G.: The measurements of observer agreement for categorical data. Biometrics 33, 159–174 (1997)
11. Laurent, D., Séguéla, P., Nêgre, S.: Cross Lingual Question Answering using QRISTAL for CLEF 2007. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2008)
12. Magnini, B., Giampiccolo, D., Forner, P., Ayache, C., Jijkoun, V., Osenova, P., Peñas, A., Rocha, P., Sacaleanu, B., Sutcliffe, R.: Overview of the CLEF 2006 Multilingual Question Answering Track. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 223–256. Springer, Heidelberg (2007)
13. Peñas, A., Rodrigo, Á., Verdejo, F.: Overview of the Answer Validation Exercise 2007. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2008)
14. Tufiş, D., Ştefănescu, D., Ion, R., Ceauşu, A.: RACAI's Question Answering System at QA@CLEF 2007. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2008)

15. Turmo, J., Comas, P., Ayache, C., Mostefa, D., Rosset, S., Lamel, L.: Overview of QAST 2007. In: Peters, C., et al. (eds.) CLEF 2007. LNCS, vol. 5152. Springer, Heidelberg (2008)
16. Vallin, A., Magnini, B., Giampiccolo, D., Aunimo, L., Ayache, C., Osenova, P., Peñas, A., de Rijke, M., Sacaleanu, B., Santos, D., Sutcliffe, R.: Overview of the CLEF 2005 Multilingual Question Answering Track. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 307–331. Springer, Heidelberg (2006)
17. Voorhees, E.: Overview of the TREC 2002 Question Answering Track. In: The Eleventh Text REtrieval Conference (TREC 2002), National Institute of Standards and Technology, USA, NIST Special Publication 500-251 (2002)