CLEF-2004
Question Answering Track Guidelines

SUMMARY:

The Question Answering (QA) track at CLEF-2004 aims at evaluating both monolingual and cross-language systems that retrieve answers (and not entire documents) in response to natural language questions. Responses appear in large, unrestricted-domain document collections.

The goal of the CLEF QA evaluation exercise is to foster research on multilingual and cross-language systems. The track will include **seven main tasks** - each one with a specific target language - and **several sub-tasks**. Document collections and answers are formulated in the target languages, while questions may be in many source languages, according to the sub-task. The seven target languages are: Dutch, English, French, German, Italian, Portuguese and Spanish. For each of these, except English, a monolingual sub-task will be activated. Cross-language sub-tasks exploit every combination between the seven already mentioned languages. Bulgarian and Finnish have been included as additional source language, and other languages may be added, depending on participants' interest.

New tasks are similar to last year's tasks: participants will be given **200 questions** that seek short answers. Differently from the CLEF-2003 QA track, queries will ask for either **fact-based** instances or **definitions**. The latter type of question was not considered in last year's edition. Multiple instances questions (like in the TREC list task) will not be taken into consideration. Some questions may not have a known answer in the corpora, and systems should be able to recognize them.

Document collections will be distributed to registered participants via the CLEF website, while question sets will be released on the 10th of May 2004. Participants will have one week to process the data, so results are due by the 17th of May.
No manual intervention of any kind is allowed: participants must freeze their systems before downloading the questions. Systems must return **exactly one response per question**, and **up to two runs**. In each run, the question set must be returned unranked, in the same order as it will be released.

Human assessors will manually check **responsiveness** (considering each `[answer-string, document-id]` pair) and **exactness** (considering each answer-string) of the answers, and individual results will be released starting from the 15th of July.
Each `[answer-string, document-id]` pair will be assessed and marked with one of the following judgments: wrong (W), unsupported (U), inexact (X) or right (R). Factoid and definition questions will not be assessed separately, and they will be given the same weight in the total score.

All judgments are according to the opinion of human assessors (whose mother tongue is the target language of the task). The total score for each run will be calculated computing **accuracy** (i.e. the fraction of correct answers).
As an additional score, that will be assigned only to those systems that can express their confidence in responses, **confidence weighted measure** will be computed.

DOCUMENT COLLECTIONS:

Registered participants can download the corpora from the CLEF website (registration form and end-user agreement must be first filled in).
The document set for the CLEF-2004 QA track is larger than the one that was used last year: new corpora for 1995 have been added.
The texts are SGML tagged, and encoded in ISO Latin-1 (ISO-8859-1). Corpora are made up of newspaper and news agency articles, and each document has a unique `document-id` identifier, that systems must return together with the answer.

Here is a sample from the Spanish EFE document collection:

```
<DOC>
<DOCNO>EFE19940101-00001</DOCNO>
<DOCID> EFE19940101-00001</DOCID>
<DATE>19940101</DATE>
<TIME>00.28</TIME>
<SCATE>POX</SCATE>
<FICHEROS>94F.JPG</FICHEROS>
<DESTINO>ICX EXG</DESTINO>
<CATEGORY>POLITICA</CATEGORY>
<CLAVE>DP2403</CLAVE>
<NUM>736</NUM>
<PRIORIDAD>U</PRIORIDAD>
<TITLE> GUINEA-OBIANG
        PRESIDENTE SUGIERE RECHAZARA AYUDA EXTERIOR CONDICIONADA
              ...
</TITLE>
<TEXT>  Malabo, 31 dic (EFE). El presidente de Guinea Ecuatorial,
 Teodoro Obiang Nguema, sugirió hoy, viernes, que su Gobierno
 podría rechazar la ayuda internacional que recibe si ésta se
 condiciona a que en el país haya "convulsiones políticas".
 En su discurso de fin de año,
              ...
 ... conceptos de libertad, seguridad ciudadana y
 desarrollo económico y social. EFE
  DN/FMR
  01/01/00-28/94
 </TEXT>
 </DOC>
```

The following table explains what text collections will be used for each target language:

TARGET LANGUAGE	CORPUS	YEAR	SIZE
Dutch	NRC Handelsblad	1994 1995	299 MB (84,121 docs)
	Algemeen Dagblad	1994 1995	241 MB (106,483 docs)
English	Los Angeles Times	1994	425 MB (113,005 docs)
	Glasgow Herald	1995	154 MB (56,472 docs)
French	Le Monde	1994	157 MB (44,013 docs)
	SDA French	1994	86 MB (43,178 docs)
	SDA French	1995	88 MB (42,615 docs)
German	Frankfurter Rundschau	1994	320 MB (139,715 docs)
	Der Spiegel	1994 1995	63 MB (13,979 docs)
	SDA German	1994	144 MB (71,677 docs)
	SDA German	1995	141 MB (69,438 docs)
Italian	La Stampa	1994	193 MB (58,051 docs)
	SDA Italian	1994	85 MB (50,527 docs)
	SDA Italian	1995	85 MB (48,980 docs)
Portuguese	PÚBLICO	1994 1995	348 MB (106,821 docs)
Spanish	EFE	1994	509 MB (215,738 docs)
	EFE	1995	577 MB (238,307 docs)

QUESTIONS:

In all tasks, systems will receive **200 short and fact-based questions** as input.
Differently from CLEF-2003, where only **factoid** queries were proposed, a small number of **definitional** questions (for instance "*What is ESCWA?*") will be included, as well.
Other categories of queries, such as 'subjective' questions (e.g. "*What is the most beautiful city in the world?*"), multiple instance questions (e.g. "*List modernist novelists.*"), 'nested' questions (e.g. "*When did the king who succeeded Queen Victoria die?*"), 'closed' questions (e.g. "*Did Shakespeare write any sonnets?*") and Why- questions (e.g. "*Why did global warming take hold as world concern?*") will not be taken into consideration.
Questions may have many answer types: the date of an event, the measure of a distance or duration, a location, a person, an organization, the manner in which an event occurs (as in How- questions), an abstraction, etc.

Some questions may even have no answer in the document collection, and in this case the correct response is a blank string with document-id NIL. A question is assumed to have no right answer when neither human assessors nor participating systems can find one.

As far as **definition questions** are concerned, they will address exclusively **people** (a person's job, position, role, etc.) and **organizations** (their name, activity, commitment, etc.), as in "*Who is G. W. Bush?*" and "*What is UNICEF?*".
Due to the difficulties encountered at TREC-2003, and considering the reliability of the "information nuggets" approach that was used to assess them, "concept definition questions" (e.g. "*What is communism?*") will be avoided. Being asked in isolation, they can have many answers, more or less detailed, and it is difficult to establish what is the most useful and objective response for them.
As a first step in the discussion on definition questions, some of them (**about 20**) will be included in each test set. They will deal with persons and organizations, assuming that the questioner does not know anything about them. Potential users of a QA system may come across an unknown person or organization, and the answer should offer them some fundamental information, in order to give a general frame that would be **useful** to understand the article they are reading.
The definition question "*What is UNICEF?*" could have both "*United Nations Children's Fund*" and "*UNICEF looks after the needs of children and mothers in developing countries around the world*" as right answers, while the answer-string "*UNICEF was created in December 1946 by the United Nations*" would not be responsive.

Question sets for all tasks will be released on the 10th of May, and participants will have one week to process them and return their results.
Test sets will be formatted as plain text files (UTF-8 encoded), with one question on each line. Lines will be structured as follows:

	task identifier			
question type	source language	target language	question number	string
F|D	BG|DE|EN|ES|FI|FR|IT|NL|PT	DE|EN|ES|FR|IT|NL|PT	4 digits	question

Question type is described in the first column, where F stands for factoid and D for definition.
Languages are described by an abbreviation (BG=Bulgarian, DE=German, EN=English, ES=Spanish, FI=Finnish, FR=French, IT=Italian, NL=Dutch and PT=Portuguese) and question number ranges between 0001 and 0200.
So, each file is 200 lines long and each column in a single line is separated by a blank space.

For instance, the question:

```
F IT EN 0092 Dove si trova il Centro Pompidou?
```

would be the ninety-second one in the test set for the cross-language task IT=>EN, where Italian queries search for an answer in an English document collection.

ANSWERS:

Participants will have one week to process the data, and responses are due by the 17th of May. The submission procedure has not yet been defined.
While last year's participants were allowed to submit up to three answers, this year they must return exactly one answer per question, and up to two runs.
An answer is basically structured as an [answer-string, document-id] pair, where the answer-string contains nothing more than a complete and exact answer (the minimum information required) and the document-id is the unique identifier of a document that supports the answer. There are no particular restrictions on the length of an answer-string (which is normally very short), but unnecessary pieces of information will be penalized, since the answer will be marked as non-exact. Because definition questions may have long strings as answers, assessors will be less demanding in judging their exactness: assessors will mainly focus on their responsiveness and usefulness.
Each submitted run is a single file, that contains exactly 200 lines (one line per question). Each line is in the following format:

question type	1 ≤ n ≤ 200	run-tag	0 ≤ n ≤ 1	document-id	answer-string
F	92	irst041iten	0.257	LAT19940311.00318	Paris

where:

- the **first column** is the question type, that can be either F (factoid) or D (definition), as in the test set.
- the **second column** is the question number: questions must be returned in the same ascending (increasing) order in which they appear in the test set, i.e. from 1 to 200.
- the **third column** is the run-tag, that describes:
  - the name of the participating group (arbitrary sequence of four ASCII characters),
  - the current year (04 stands for 2004),
  - the number of the run (1 if it is the first one, or 2 if it is the second one),
  - the task identifier (the same of the test set)

Clearly, the content of this field never changes within the same submission file. Each submission file must be named after this column, with a .txt extension. So, the line in the example above would be included in the file `irst041iten.txt`.

- the **fourth column** is the confidence number, that is a mandatory integer or floating point value (maximum length is 8 characters) that can range between `0` and `1`, inclusive, where 0 means that the system has no evidence of the correctness of the answer, and 1 means that the system is absolutely confident about the correctness of the answer.
  If a system returns integer or floating point confidence values that are higher than 1, it must normalize them for each response. If a system does not produce any score number, it must return a default score equal to 0 (zero).
  Score value will be used in a second, additional evaluation (the main measure is accuracy) in order to test systems' self-evaluation ability.
- the **fifth column** contains the `document-id` of the document that supports the given answer. Some questions may not have a known response in the document collection: in that case the correct answer would be a blank answer-string, and the string `NIL` would replace the document-id identifier (see fifth line in the sample below).
- the **sixth column** contains the answer-string, that is blank if the document-id is NIL. There can be no line breaks in the answer-string.
  Human assessors will judge both correctness and exactness of the submitted answers, so unnecessary additional information will be marked as non-exact. The same judgment will be assigned to responses that lack significant bits of information.

Generally speaking, participants should follow the *all columns must be present* rule, except when the fifth column is NIL. There must be at least one blank space between columns, but lines must not be longer than 1024 bytes. There should be a single line-break after each answer string, so that the next answer starts on the very next line. In order to facilitate participants, a checking routine for the submissions will be released.

Here is an example of how the first lines of a submission file might look:

```
F   1 irst041iten 0.005 LAT19941109.01011 oil
F   2 irst041iten 0.343 GH19950230.00188 perhaps a thousand years old
F   3 irst041iten     1 LAT19940122.00022 yellow
F   4 irst041iten 0.201 LAT19940327.00198 Kennedy
D   5 irst041iten 0.012 NIL
D   6 irst041iten 0.207 GH19951007.00036 30-storey building
F   7 irst041iten 0.802 LAT19940913.00337 a tasty apple pie
```

EVALUATION:

The files submitted by participants in all tasks, where each task corresponds to a target language, will be manually judged by native speaking assessors. Assessors will consider

correctness (i.e. responsiveness) and **exactness** (i.e. the quantity of information) of the returned answers. They will also check that the document labelled with the returned document-id supports the response.

Each `[answer-string, document-id]` pair will be assessed and marked with one of the following judgments, that are used at TREC, as well:

- **incorrect**: the answer-string does not contain a correct answer or the answer is not responsive;
- **unsupported**: the answer-string contains a correct answer but the returned document id does not support it;
- **non-exact**: the answer-string contains a correct answer and the returned document id supports it, but the string is missing bits of the answer or is longer than just the minimum exact answer;
- **correct**: the answer-string consists of exactly a correct answer, and the answer is supported by the returned document.

Individual results will be released to each participating group from the 15th of July. All runs will be judged, and assessors will attach their judgements at the beginning of each line returned by the system, as follows:

```
W F   1 irst041iten 0.005 LAT19941109.01011 oil
X F   2 irst041iten 0.343 GH19950230.00188 perhaps a thousand years old
R F   3 irst041iten     1 LAT19940122.00022 yellow
U F   4 irst041iten 0.201 LAT19940327.00198 Kennedy
W D   5 irst041iten 0.012 NIL
W D   6 irst041iten 0.207 GH19951007.00036 30-storey building
X F   7 irst041iten 0.802 LAT19940913.00337 a tasty apple pie
```

where:

`W` stands for wrong,
`U` stands for unsupported,
`X` stands for inexact,
`R` stands for right.

EVALUATION MEASURE:

In last year's CLEF QA track participants were allowed to submit up to three answers per question (sorted by confidence), and the score assigned to each question was the reciprocal of the rank for the first response to be judged correct. The total score was calculated computing the Mean Reciprocal Rank.

This year systems must return exactly one response per question. Each run is an unranked list of answers, sorted in the same order as the test set. **The main evaluation score of a run is accuracy**, i.e. the fraction of right answers (R).

Since systems return a **confidence** value for each answer, a **second measure** will be computed.

Each question will score 1 if its answer is judged correct, and 0 otherwise. Then answers will be sorted according to their confidence score, and the total score of a run will be determined using the evaluation measure that in the TREC-2002 QA track guidelines is described as "*analogue to document retrieval's uninterpolated average precision*". According to this measure, which rewards the systems that can evaluate their own performance, the final score ranges between 0 and 1, inclusive, with 1 being a perfect score. This additional measure is computed as:

```
 sum for i=1 to 200 (#-correct-up-to-question-i/i)
---------------------------------------------------
                         200
```

where 200 is the number of the questions in all tasks, and i ranges over the questions.

Though mandatory, some systems may not be able to return a confidence value different from 0: in that case confidence weighted score will not be computed.

SCHEDULE:

Registration Opens	January 15, 2004
Corpora Release	February 2004
Trial Data	March 2004
Test Sets Release	May 10, 2004
Submission of Runs by Participants	May 17, 2004
Release of Individual Results	from July 15, 2004
Submission of Papers for Working Notes	August 15, 2004
CLEF Workshop	15-16 September, 2004

Appendix:

CORRECTNESS AND EXACTNESS

Although it is difficult to decide a priori what are the criteria that determine exactness and correctness of an answer, assessors are given the following set of tentative rules according to which they will judge the submissions. Different answer types seem to have different requirements.

Anyhow, each answer will be examined considering the particular question it is referred to, and judgments will depend on contextual cases.

Location:

Assessors will judge the answers considering the conventional named entities that appear in atlases. For instance, in Italian some locations require a specific name (e.g. "*Mar Nero*", and not only "*Nero*", which would be incorrect), while other can have many acceptable forms (e.g. "*Mar Tirreno*" and also "*Tirreno*").

In response to some questions, country-only designation can be judged as sufficient, but normally queries seek specific and precise answers. If a document reported that "*the coldest place in the world is Plateau Station, Antarctica*", "*in the world*" would not be an acceptable response to the question "*Where is Plateau Station*?".

Time:

As regards the date of specific concluded events, day/month/year or month/year specifications are normally required, but not month/day or day only (unless the question explicitly asks for one of them). If day and month cannot be retrieved, the year is sufficient.

For example, if a system answered the question "*When did Napoleon die?*" returning the string "*5th of May*", such answer would be marked as incorrect or non-exact. On the other hand, both "*5th of May 1821*" and "*1821*" would be correct and exact answers.

Some questions may refer to recurring events, such as anniversaries and birthdays. In case of anniversaries month/day designation would be considered sufficient: "*4th of July*" would be a correct string in response to the question "*When is Independence Day?*". As far as birthdays are concerned, those of living persons need not include the year but those of dead people usually do.

Relative answers could be accepted, as well. A question like "*When was Nicole Kidman born?*" could have both "*on 20 June 1967*" (absolute answer) and "*36 years ago*" (relative answer) as correct responses. Clearly, the latter should be supported by a document, and not computed by the system. In addition, relative answers can be accepted only if they are responsive: "*two years before her brother*" would be judged as wrong.

Person:

As a general rule, first name only would be regarded as insufficient, while last name only is probably sufficient.

Nevertheless, if in a document the same last name refers to more than one person, assessors need the first name to decide how to judge the answer. So, assessors will take into consideration the context of the entire document.

Sometimes, questions may have more than one answer type, like "*Who adopted a new constitution in Africa?*". Both "*Nelson Mandela*" (person) and "*South African government*" (organization) could be correct and exact answers.

Measure:
Quantitative answers are considered responsive when they include units. For example, with the question "*What is the world's population?*", an answer like "*5.5*" would be incorrect, because it lacks the unit "*billion*" or "*bn*". In addition, punctuation is necessary: "*5 5 billion*" would not be acceptable.
Even though a document stated that "*world's population is six times the population of China*", the answer "*six times the population of China*" would express a comparison, without providing any real quantitative estimate, and thus it would be incorrect.
Sometimes units are not strictly necessary: "*32,186*" would be a sufficient and correct string in response to the question "*How many meters are in 20 miles?*".

Manner:
Test sets could contain How- questions, like "*How did Ayrton Senna die?*". Since they can have many possible answer types, these questions are similar to definition questions. They can have many acceptable answers, that can be more or less related to the event that the question hints at. So, assessors could judge as correct both the answers "*in a car accident*" and "*tragically*".

Replicas:
Following the TREC QA guidelines, we can assume that, unless a question specifically states otherwise, any question regarding a famous entity is asking about the famous entity and not about copies or imitations.

Articles, prepositions and other parts of speech:
As far as the exactness of an answer is concerned, articles and prepositions might be acceptable parts of speech that can be included in an exact answer-string. On the contrary, other parts of speech such as adjectives and adverbs could add subjective or unnecessary information.
For instance, consider the question "*In what year did Napoleon die?*". Answers like "*1821*", "*in 1821*" or "*in the year 1821*" would be judged as exact, while "*in the terrible year 1821*" would be non-exact.
Anyhow, it is difficult to formulate a priori hypotheses on exactness, and judgments will be in the opinion of human assessors who, whenever they have doubts, will try to reach an agreement. Definition questions may require long circumlocutions as answers, and in that case adjectives and adverbs could be important.

Particular cases:
If a document reports wrong information (for example that Napoleon died in 1831), and a system returns this wrong information (for example "*in 1831*") with the correct document-id, the answer is judged correct.

If the requested answer-string appears many times in the same document but in different contexts (for example the document could say in different passages that Napoleon in 1821 wrote his diary, was visited by a friend and died) the answer-string (e.g. "*1821*") supported by that document is considered correct.