QA@CLEF-2005

Guidelines

DOCUMENT COLLECTIONS

Registered participants can download the corpora from the CLEF website (registration form and end-user agreement must be first filled in).
The table below shows which document collections will be used for each target language.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
BG	x	x															
DE										x	x	x					
EN					x	x											
ES															x		
FI							x										
FR								x	x								
IT													x	x			
NL			x	x													
PT																x	x

Where:

1. Sega 2002
2. Standart 2002
3. NRC Handelsblad 1994 + 1995
4. Algemeen Dagblad 1994 + 1995
5. Los Angeles Times 1994
6. Glasgow Herald 1995
7. Aamulehti 1994 + 1995
8. Le Monde 1994
9. French SDA 1994 + 1995
10. Frankfurter Rundschau 1994
11. Der Spiegel 1994 + 1995
12. German SDA 1994 + 1995
13. La Stampa 1994
14. Italian SDA 1994 + 1995

15. EFE 1994 + 1995
16. Público 1994 + 1995
17. Folha 1994 + 1995

Note: Le Monde 1995 (that was used last year) will not be used this year.

QUESTIONS

This year we will propose three types of questions:

1. temporally unrestricted factoids (F)
2. temporally restricted factoids (T)
3. definition questions (D)

1. These are the usual factoid questions we proposed last year, i.e. questions that address locations, persons, measures, etc.
2. We assume three types of temporal restriction for factoids:
   - by **date** (e.g. "Who was the president of the United States in 1910?")
   - by **period** (e.g. "Who was the president of the United States between 1910 and 1917?")
   - by **event** (e.g. "Who was the president of the United States when the Berlin Wall was torn down?")

   Similar questions were proposed last year at the Spanish pilot QA track. You can find more examples here.

3. These questions will address exclusively **organizations** (their name, activity, committment, etc.), e.g. "What is Amtrack?" and **people** (a person's job, position, role, etc.), e.g. "Who was Marilyn Monroe?".

Test sets will be made up of 200 questions. Most of them will be temporally unrestricted factoids. There will be around 50 definitions and 30 temporally restricted factoids.
Some questions may even have **no answer** in the document collection, and in this case the correct response is a blank string with docid "NIL". A question is assumed to have no right answer when neither human assessors nor participating systems can find one.
This year neither How- questions, nor Yes/no questions, nor List questions will be proposed.
Question sets for all tasks will be released on the 18th 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 per line.
Each line will be structured as follows:

task identifier

question type	source language	target language	question number	UTF-8 encoded string
F|D|T	BG|DE|EN|ES|FI|FR|IN|IT|NL|PT	BG|DE|EN|ES|FI|FR|IT|NL|PT	4 digits	question

Question type is described in the first column, where `F` stands for factoids, `D` for definitions and `T` for temporally restricted factoids.
Question number ranges between `0001` and `0200`.
The task is described in the second and third column.
Each file is 200 lines long and columns are separated by a single blank space.
For instance, the first three questions in the EN-ES test set (English questions that hit a Spanish document collection) might look like these:

F	EN	ES	0001	In which country is the Cape of Good Hope?
D	EN	ES	0002	What is Amtrack?
T	EN	ES	0003	Who was the president of the United States in 1961?

As you can see, format has not changed since last year.

IMPORTANT UPDATE: May 19, 2005
The order of the columns in the template above is different from the one in the test sets released for CLEF-QA 2005, i.e. the fourth column in the template is the second one in the test set, so that the actual format in the test set is as in the following example:

F	0001	EN	ES	In which country is the Cape of Good Hope?

ANSWERS

Each participating group will be allowed to participate in any task. Anyhow, we encourage participants (especially veterans) to consider questions and target languages other than their own language and English.
Participants will have **one week** to process the data, and responses are due by the 25th of May. Instructions concerning the results submission procedure will be given when the track will start.
As last year, participating teams must return **exactly one answer per question**, and **up to two runs**.
Systems must give an answer to all the questions. Partial submissions will not be accepted.
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	question number	team + year + run identifier + task identifier	confidence score	docid	answer-string

Examples:

F	0001	irst051enes	0.861	EFE19940427-16057	Sudáfrica
D	0002	irst051enes	0.113	NIL	
T	0003	irst051enes	0.669	EFE19940118-08606	John Fitzgerald Kennedy

where:

- **the first column** is the question type, that can be either `F` (factoid), `D` (definition) or `T` (temporally restricted factoid), as in the test set.
- **the second column** is the question number: answers must be returned in the same ascending (increasing) order in which questions appear in the test set, i.e. from `0001` to `0200`.
- **the third column** describes:
    - the name of the participating team (arbitrary sequence of four ASCII characters),
    - the current year (`05` stands for 2005)

      ,

    - the identifier of the run (`1` if it is the first one, or `2` if it is the second one),
    - the task identifier (including both source and target languages, as in 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 lines in the examples above would be part of the file `irst051enes.txt`.

the fourth column is the confidence score, 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 docid of the document that supports the given answer. Some questions may not have any 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 docid (see second line in the examples above).

the sixth column contains the answer-string, that is blank if the docid 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 has been released (see page resources). We strongly encourage participants to run the script and check their submissions before uploading them.

EVALUATION

The files submitted by participants in all tasks 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 docid supports the given response.
Each [answer-string, docid] 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 docid does not support it;
- **non-exact**: the answer-string contains a correct answer and the docid 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.

More details concerning correctness and exactness can be found in last year's guidelines.
Individual results will be released to each participating group from the 18th of July. All runs will be judged, and assessors will attach their judgments at the beginning of each line returned by the system, as follows:

```
W F  0001 irst052iten 0.005 LAT19941109.01011 oil
X F  0002 irst052iten 0.343 GH19950230.00188 perhaps a thousand years old
R F  0003 irst052iten 1 LAT19940122.00022 yellow
U F  0004 irst052iten 0.201 LAT19940327.00198 Kennedy
W D  0005 irst052iten 0.012 NIL
W T  0006 irst052iten 0.207 GH19951007.00036 30-storey building
X F  0007 irst052iten 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 MEASURES:

The main evaluation score of a run is **accuracy**, i.e. the fraction of right answers (R).
In addition, the **correlation between correctness of the answers and system self-scoring** will be explored.

SCHEDULE

Test set release	May 18, 2005 (12 am CET)
Submission of runs by participants	by May 25, 2005 (12 am CET)
Release of individual results	July 18, 2005
Submission of papers for the CLEF Working Notes	by August 21, 2005
CLEF workshop	21-23 September, 2005 (in Vienna, Austria)