GUIDELINES for the PARTICIPANTS in QA@CLEF 2008

INTRODUCTION

The main task scenario is event-targeted QA on a heterogeneous document collection (news articles and Wikipedia). Many monolingual and cross-language sub-tasks are offered: Bulgarian, English, French, German, Italian, Portuguese, Romanian, Greek, Basque and Spanish are proposed as both query and target languages.

A large number of questions will be topic-related, i.e. clusters of questions which are related to the same topic and possibly contain anaphoric references between one question and the other questions.

Besides the usual news collections, articles from Wikipedia will be considered as an answer source. Some questions may have answers only in one collection, i.e. only in the news corpus or in Wikipedia.

ORGANIZING COMMITTEE

- Anselmo Peñas (UNED, Spain) Co-chair
- Danilo Giampiccolo (CELCT, Italy) Co-chair
- Corina Forascu (Romania Institute for Computer Science, Romania)
- Gosse Bouma (University of Groningen)
- Nicolas Moreau (ELDA/ELRA, France)
- Pamela Forner (CELCT, Italy)
- Petya Osenova (BTB, Bulgaria)
- Paulo Rocha (Linguateca, DEI UC, Portugal)
- Bogdan Sacaleanu (DFKI, Germany)
- Richard Sutcliffe (University of Limerick, Ireland)
- Prokopis Prokopidis (ILSP Athena Research Center, Greece)
- Iñaki Alegria (University of Basque Country, Spain)

IMPORTANT DATES

Registration at the website[1] (participants must communicate the tasks (language pairs) in which they are going to participate): **March 31st**

Question sets for all tasks will be released on the **May 19th**.

Runs must be returned within 5 days from the first test set download no later than **May 27 th**.

[1] http://nlp.uned.es/clef-qa (Please note that you must also register at the Clef Campaign website http://www.clef-campaign.org/)

Instructions concerning the results submission procedure will be given when the track will start.

Individual results will be released to each participating group from **July 1st**.

DOCUMENT COLLECTIONS

Registered participants can download the corpora from the CLEF website (registration form and end-user agreement must be first filled in).

Target document collections are the following:

1. Newswire Collections

TARGET LANGUAGE	COLLECTION	PERIOD
Basque (EU)	Egunkaria	2001/2003
Bulgarian (BG)	Sega	2002
	Standart	2002
	Novinar	2002
Dutch (NL)	NRC Handelsblad	1994/1995
	Algemeen Dagblad	1994/1995
English (EN)	Los Angeles Times	1994
	Glasgow Herald	1995
French (FR)	Le Monde	1994
	Le Monde	1995
	French SDA	1994
	French SDA	1995
Germany (DE)	Frankfurter Rundschau	1994
	Der Spiegel	1994/1995
	German SDA	1994
	German SDA	1995
Italian (IT)	La Stampa	1994
	Italian SDA	1994
	Italian SDA	1995
Portuguese (PT)	Público	1994
	Público	1995
	Folha de São Paulo	1994
	Folha de São Paulo	1995
Spanish (ES)	EFE	1994
	EFE	1995
Greek (EL)	The Southeast European Times	2002

2. In addition, the **Wikipedia pages dumped at November 2006** must be used. They can be used in any format (XML or HTML versions). Since answers could be inserted in tables or any other html element, the use of the HTML format (or other format preserving this information) is strongly recommended. In this way, the main reference for the submission assessments will be the original pages visualized with usual web browsers. At least the HTML versions for each target language will be downloadable under the restricted area at http://nlp.uned.es/clef-qa after login. Other formats can be provided if they are available.

Only pages from this "snapshot" of Wikipedia November 2006 will be allowed since it is impossible to prevent pages or topics from changing in the online version of Wikipedia.

All Wikipedia answers must be taken from "actual entries" or articles of Wikipedia pages - the ones whose filenames normally correspond to the topic of the article. Other types of data ("image", "discussion", "category", "template", "revision histories", any files with user information, and any "meta-information" pages), must be excluded.

QUESTIONS

Following last year experience, the exercise will consist of **topic-related questions**, which means clusters of questions which are related to the same topic and possibly contain co-references between one question and the others.

Topics can 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 is related to the topic with the following structure:

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

Anyway, topics will NOT be given in the test set but may be inferred from the first question/answer pair.

For example:

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

Or

TOPIC: George W. Bush
Q1: Who was the President of the United States in 2002?
Q2: When was he born?
Q3: Who was his wife?

As far as the question types are concerned, as in previous years of QA@CLEF, we still consider the following three question categories:

a) factoid
b) definition
c) closed list

a) Factoid questions are 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. **Q:** *Who was called the "Iron-Chancellor"?*
  A: *Otto von Bismarck.*
- TIME, e.g. **Q:** *What year was Martin Luther King murdered?*
  A: *1968.*
- LOCATION, e.g. **Q:** *Which town was Wolfgang Amadeus Mozart born in*?
  A: *Salzburg.*
- ORGANIZATION, e.g. **Q:** *What party does Tony Blair belong to?*
  A: *Labour Party.*
- MEASURE, e.g. **Q:** *How high is Kanchenjunga?*
  A: *8598m.*
- COUNT, e.g. **Q:** *How many people died during the Terror of Pol Pot?*
  A: *1 million.*
- OBJECT, e.g. **Q:** *What does magma consist of?*
  A: *Molten rock.*
- OTHER, i.e. everything that does not fit into the other categories above.
  Q: *Which treaty was signed in 1979?*
  A: *Israel-Egyptian peace treaty.*

b) Definition questions are 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, **Q:** *Who is Robert Altmann?*
  A: *Film maker.*
- ORGANIZATION, i.e. questions asking for the mission/full name/important information about an organization, e.g.
  Q: *What is the Knesset?*

A: *Parliament of Israel.*

- OBJECT, i.e. questions asking for the description/function of objects, e.g.
  Q: *What is Atlantis?*
  A: *Space Shuttle.*

- OTHER, i.e. question asking for the description of natural phenomena, technologies, legal procedures etc., e.g.
  Q: *What is Eurovision?*
  A: *Song contest.*

c) Closed list questions: i.e. questions that require in **one** single answer the requested number of items, e.g:

Q: *Name all the airports in London, England.*
A: *Gatwick, Stansted, Heathrow, Luton and City.*

Q: *Name the last three American Presidents.*
A: *George H.W. Bush, Bill Clinton, George W. Bush.*

All types of questions may contain a **temporal restriction**, i.e. a temporal specification that provides important information for the retrieval of the correct answer. Examples:

Q: *Who was the Chancellor of Germany from 1974 to 1982?*
A: *Helmut Schmidt.*
Q: *Which book was published by George Orwell in 1945?*
A: *Animal Farm.*
Q: *Which organization did Shimon Perez chair after Isaac Rabin's death?*
A: *Labour Party Central Committee.*

Test sets will be made up of 200 questions, most of which will be temporally unrestricted factoids.

Some questions may even have no answer in the document collection, and in this case the exact answer is "NIL" and the answer and support docid fields are empty. A question is assumed to have no right answer when neither human assessors nor participating systems can find one.

NB: the question type will not be provided to the systems.

IMPORTANT: The **NOW** of a question (and of its answer) may be problematic, as different document collections from different time spans are used. As a consequence, the temporal collocation of the query must be understood as the one indicated in the document from which the exact answer is retrieved.

ANSWERS

Each participating group will be allowed to participate in any task. We encourage participants (especially "veterans") to consider questions and target languages other than their own language and English.

Participating teams must return **up to three answers per question**, and **up to two runs.** All questions must be answered and no partial submissions will be accepted.

Each exact answer must be supported by:

- the DOCNO of the document in the news collection (not the DOCID which is different in certain collections, unless DOCNO is not provided) or by the filename of the dumped Wikipedia (November 2006) page, from which it has been retrieved. In either case leading and trailing spaces should be removed. In consequence, DOCNOs for newspaper articles will contain no white space at all (e.g. "LA112994-0248) while the identifier of a Wikipedia article could contain white spaces exactly as it was originally, but with no leading or trailing spaces (e.g. "Nights in the gardens of Spain");
- portion(s) of text, which provide enough context to support the correctness of the exact answer. Supporting texts may be taken from different sections of the relevant document, and which must sum up to a maximum of 700 bytes. Unnecessarily long snippets, i.e. those that do not meet this requirement, might be judged as non-supporting.

  All the information needed to support the exact answer should be put in the supporting text field, and not simply mentioned in the ID/FILENAME. This is vital as regards answers taken from Wikipedia, where the topic is usually mentioned only in the title, and then referred to only anaphorically in the actual text. Snippet should contain the title of the page or something else to solve the co-reference.
  Systems should find the way to give all the supporting information in the snippet.

  The same holds for the date issue. All the information (either date or topic) should be explicitly mentioned in the supporting snippets.

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. The answer string must contain nothing more than a complete and exact answer, i.e. the minimum amount of information needed to satisfy the query.

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.

As in previous years, the exact answer may be copied/pasted from the document even if it is grammatically incorrect (e.g.: inflectional case does not match the one required by the question). In addition, systems will be allowed to use NL generation in order to correct morpho-syntactical inconsistencies (e.g., in German, changing "dem Präsidenten" into "der Präsident" if the question implies that the answer is in Nominative case), and to introduce grammatical and lexical changes (e.g., Q: What nationality is X? TEXT: X is from the Netherlands => EXACT ANSWER <Dutch>).

INPUT FORMAT

DTD for the INPUT format can be downloaded from the web site (http://nlp.uned.es/clef-qa/)

Test sets will be formatted as an xml file (UTF-8 encoded).
The xml will be structured with elements containing the following information:

- Source language <*source_lang*>
- Target language <*target_lang*>
- Question group id (4 digits – 1000 to n) <*q_group_id*>
- Question number (4 digits – 0001 to 0200) <*q_id*>
- Question (UTF-8 encoded string) <*q_string*>

i.e.:

```
<?xml version="1.0" encoding="UTF-8" ?>
<input>
      <q q_id="0001-0200" q_group_id="1000-n"
      source_lang="BG|DE|EL|EN|ES|EU|FR|IT|NL|PT|RO"
      target_lang="BG|DE|EL|EN|ES|EU|FR|IT|NL|PT|RO">Question?</q>
</input>
```

Example:
The first three questions in the EN-ES test set – i.e. English questions that hit a Spanish document collection - might be represented as follows:

```
<?xml version="1.0" encoding="UTF-8" ?>
<input>
      <q q_id="0001" q_group_id="1000" source_lang="EN" target_lang="ES">
      What is a blackberry?</q>
      <q q_id="0002" q_group_id="1001" source_lang="EN"
      target_lang="ES">In which country is the Cape of Good Hope?</q>
      <q q_id="0003" q_group_id="1001" source_lang="EN" target_lang="ES">
      What is its capital city?</q>
</input>
```

N.B.: The second and third questions above belong to the same question group (id 1001); question 0003 contains a co-reference to the previous answer, which must be resolved in order to retrieve the answer.

OUTPUT FORMAT

DTD for the OUTPUT format can be downloaded from the web site (http://nlp.uned.es/clef-qa/)

The format of each system's output must conform to the following xml (**UTF-8 encoded**) with each line containing one answer:

```
<?xml version="1.0" encoding="UTF-8" ?>
<output>
      <a q_id="0001-0200" q_group_id="1000-n" run_id="XXXX08XXXX"
      score="0.nnn+">
      <answer>xyz</answer>
      <docid>docid |Wikipedia HTML filename| Wikipedia XML
      title</docid>
      <support>
      <s_id>docid |Wikipedia HTML filename| Wikipedia XML
      title</s_id>
      <s_string>xyz.</s_string>
      </support>
      </a>
</output>
```

Where:

- q_group_id : the question group number as appears in the test set, i.e. from 1000 to n, which in the input file identifies a group of questions referring to the same topic.
- q_id : the question number as given in the test set. Answers must be returned in the same ascending (increasing) order in which questions appear in the test set, i.e. from 0001 to 0200.
- run_id : an alphanumeric string which identifies the runs of each participant. It describes, in one single string:
    - the **name of the participating team** (sequence of four lower case ASCII characters)
    - the **current year** (08 stands for 2008)
    - the **number 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 all in lower case, with a .txt extension, e.g. clct081itit.txt.

- score (confidence score): a mandatory floating point value (maximum length is 8 characters) that can range between 0.0 and 1.0, inclusive, where 0.0 means that the system has no evidence of the correctness of the answer, and 1.0 means that the system is absolutely confident about the correctness of the answer. Values **must** be normalized to the range 0.0 ↔ 1.0. If a system does not produce any score number, it must return a default score equal to 0.0. Score value will be used in a second, additional evaluation (the main measure is accuracy) in order to test systems' self-evaluation ability.

- and the <a> element has the following children elements:
  - docid: the answer docno/Wikipedia HTML filename/Wikipedia XML title element contents that supports the exact answer. Some questions may not have any known response in the document collection: in that case the docid element remains empty.
  - answer : contains the exact answer-string that is NIL if no answer has been retrieved in the document collections. In such case the <docid>; <s_id> ; and <s_string> elements are empty.
  - s_id : the supporting text docid/ Wikipedia HTML filename/Wikipedia XML title element contents that identifies the file from which the supporting text is taken.
  - s_string: the supporting text. Each answer must have at least one supporting snippet, except the NIL answers. Supporting texts may be taken from different sections of the relevant document, in which case they will be located in different tagged elements up to a maximum of three. The total length of the supporting texts must not exceed 700 bytes.

Example:

```
<?xml version="1.0" encoding="UTF-8" ?>
<output>
      <a q_id="0001" q_group_id="1000" run_id="clct081itit"
      score="0.889">
      <answer>three gold medals</answer>
      <docid>LA112994-0248</docid>
      <support><s_id>LA112994-0248</s_id><s_string>When comparing Michele
      Granger and Brian Goodell, Brian has to be the clear winner. In
      1976, while still a student at Mission Viejo High, Brian won two
      Olympic gold medals at Montreal, breaking his own world records in
      both the 400- and 1,500-meter freestyle events. He went on to win
      three gold medals in the 1979 Pan American
      Games</s_string></support>
      </a>
</output>
```

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.

Each line of the submitted runs will be assessed and marked with one of the following judgments:

- **Z** (unknown): the answer line has not been evaluated by the assessor

- **W** (incorrect): the answer-string does not contain a correct answer or the answer is not responsive;

- **U** (unsupported): the answer-string contains a correct answer but the provided text-snippets do not support it, or the snippets do not originate from the provided document.

- **X** (inexact): the answer-string contains a correct answer and the provided text-snippets support it, but the answer-string is incomplete/truncated or is longer than the minimum amount of information required;

- **R** (correct): the answer-string consists of an exact and correct answer, supported by the text snippets.

The judgments (Z/W/U/X/R) will be attached at the very beginning of each line returned by a system.

The main evaluation score of a run is **accuracy,** defined as the average of SCORE(q) over all 200 questions q, where *SCORE(q) is 1, if the first answer to q in the submission file is assessed as "*R*", and 0 otherwise.*

We will also consider another evaluation measures that use different definition of SCORE(q) to consider the three answers:

- SCORE(q) defined as the mean reciprocal rank (MRR) over N assessed answers per question. That is, the mean of the reciprocal of the rank of the first correct label over all questions. If the first correct label is ranked as the 3rd label, then the reciprocal rank (RR) is 1/3. If none of the first N responses contains a correct label, RR is 0. RR is 1 if the highest ranked label matches the correct label.

CONTACT INFORMATION

Anselmo Peñas: anselmo *at* lsi.uned.es

Danilo Giampiccolo: giampiccolo *at* celct.it

Pamela Forner: forner *at* celct.it