

## The DISEQuA Corpus

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

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

1. *Production of a pool of 200 candidate questions with their answers in each language.* These queries were formulated on the basis of the topics released by CLEF for the retrieval tasks of the year 2000, 2001 and/or 2002. The CLEF topics, i.e. a set of concepts chosen with the aim of covering the main events occurred in the years 1994 and/or 1995, allowed us to pose questions independently from the document collection. In this way we avoided any influence in the contents and in the formulation of the queries. Questions were posed according to common guidelines: they had to be generally short and fact-based, unrelated to subjective opinions. They could not ask for definitions (i.e. "Who is Bill Clinton") and they had to have just one unique and unambiguous item as response, which means that we avoided questions asking for multiple items like those used in the TREC list task. Three groups of native speakers, one for each language, were involved in this work and searched the correct answers. A question has an answer in the reference corpus if a document contains the correct response without any inference implying knowledge outside the document itself.
2. *Selection of 150 questions from each monolingual set.* Since our aim was to build a test set of shared queries that would find answers in all the monolingual corpora, each group chose 150 questions from its candidate pool and translated them into English, thus a larger collection of 450 queries was put together. English constituted a sort of inter-language we used to shift from one language to another, though we were aware that in this phase there was the risk of changing unwarily the content of the questions during the translation. Each group chose its 150 questions taking into consideration that they would be processed by the other two, so the most general queries, that were likely to find a response in the other two corpora, were selected. Those that were too strictly related to the specific issues of a country were discarded.
3. *Processing of the shared questions.* Once we had collected a pool of 450 questions that had response in one of the corpora, each group picked up the 300 questions submitted by the other two and translated them a second time from English into a new target language. As a consequence, all the questions had a translation in four different languages and could be processed again in the other two target document collections. It is important to underline that the 450 questions that form the DISEQuA corpus underwent three translations: one from the source language into English and then other two from English into the two target languages. Each translation could introduce some variations, with the risk that the four final versions would not be semantically equivalent and aligned. To avoid this problem, in the second translation both the English version and the original question in the source language were taken into consideration. When the second verification was concluded, the resulting data were merged. The different versions of the same questions were aligned, and the DISEQuA corpus was successfully assembled. Quite surprisingly, we found thirteen couples of queries that had an identical meaning, but since most of them were formulated in a slightly different way, we decided to keep them in the final version of DISEQuA. Different formulations of the same question could be exploited in Machine Translation applications. Particularly, the duplicates we found are: 5=182, 10=222, 18=274, 47=338, 50=263, 52=154, 62=169, 72=173, 79=287, 88=433, 119=203, 144=174 and 147=188.

The problem of structuring data and find a sensible format to describe both questions and answers arose during this first phase of the creation of DISEQuA. The issue was addressed conceiving an XML syntax that would show the number of each question, the keywords set (or topic) from which it was generated, the person who verified it in the document collection and the type of entity it was related to. Similarly, the answers found for each question needed to be numbered, and the docid of the document that supported each response had to be logged.

The adoption of a precise format could solve the problem of losing trace of the changes that each question could undergo, in fact new tags could be added to give more information. Secondly, structured data can be easily browsed and analysed: for instance, the tag used to indicate the question type proved to be quite useful in balancing the test set. Thirdly, a common format for questions and answers was necessary to share them between the three groups that put together the DISEQuA corpus.

## Format:

Each entry is structured in tags, attributes of the tags and values of the attributes:

The tag <qa> contains each entry of the corpus:

- the attribute `cnt` indicates the number assigned to the question (from 1 to 450);
- the attribute `type` describes the category to which the answer belongs: seven different question types were considered: PERSON, LOCATION, MEASURE, DATE, ORGANIZATION, OBJECT (i.e. concrete things) and OTHER (when the response could not be labelled with one precise type). The aim was to create a well-balanced test set, with a good coverage of all these categories.

The tag <language> describes the language of each question and answer:

- the attribute `val` indicates the language in which the question appears, so that "DUT" stands for Dutch, "ITA" for Italian, "SPA" for Spanish and "ENG" for English;
- `original` keeps track of the source and the target language of each query: This attribute can have either "TRUE" or "FALSE" as Boolean values, where "TRUE" shows that the language `val` is the source language, i.e. the language in which the question was first generated, while "FALSE" records that the query has been translated. Consequently, English questions, as intermediate versions, could have nothing but "FALSE".

The tag <question> contains the queries:

- `assessor` is an identifier of the person (or group) who processed the query, which seemed to be important in case of inconsistencies. When language `val` is "ENG", `assessor` is always empty, in fact no one processed the English questions;
- the question string was processed manually by the assessors, and then automatically by the systems that participated in the CLEF QA Track. Before generating the queries, the three groups agreed on common guidelines that would help to formulate a good and useful test set. Following the model of past TREC campaigns, and particularly of the TREC 2002 QA track, a series of basic instructions were formulated. Firstly, questions should be fact-based, and, if possible, they should address events that occurred in the years 1994 or 1995. When a precise reference to these two years lacked in the questions, it had to be considered that systems would use a document collection of that year. Secondly, questions should ask for an entity ( i.e. a person, a location, a date, a measure or a concrete object), avoiding subjective opinions or explanations. So, "Why-questions" were not allowed. Queries like "Why does Bush want to attack Iraq?" or "Who is the most important Italian politician of the twentieth century?" could not be accepted. Since the TREC 2002 question set constituted a good term of comparison, and it did not include any definitional question of the form "Who/What is X?", it was decided to avoid them, as well. Thirdly, co-ordinators agreed that multiple-item questions, like those used in the TREC list-task, should be avoided. Similarly, the people in charge for the questions generation could not formulate 'double queries', in which there is a second indirect question subsumed within the main one (for instance, "Who is the president of the poorest country in the world?"). Finally, closed questions, known as yes/no questions, should be left out, too. Queries should be closely related to the topics or to the keywords extracted from the topics, without any particular restraint in the word choice. It was not necessary to know the answer before formulating a question: on the contrary, this could influence the search for a response in the corpus.

Some candidate questions asked for events occurring "in the year 1994" (or 1995), but since 1994 (and, for Dutch, 1995) was the year in which the target corpora were published, it was very improbable that it would appear explicitly in the articles, so no document would clearly state that the year was 1994 (or 1995). For this reason, every explicit mention of the year 1994 (or 1995) had to be removed from the final version of the queries.

The tag <answer> contains the responses of each question:

- `n` represents a progressive number of responses, in fact a single query could have several correct answers in the same document collection. Dates and numbers in particular change across different news for the same event. Sometimes former news in the document collection are less precise than the latter ones, because they register a process that changes over a period. Since systems were expected to give an answer supported by a unique document, and not the final or best answer in the whole corpus, in such cases there were many correct responses.
- the attribute `idx` gives the docid identifier of the document in which each single answer appears. Systems should return the docid as a justification of the answer, and in strict evaluation the unsupported responses were considered as incorrect.

When no answer was found in the target corpus, answer `n` and `idx` were labelled with 0 (zero), and the answer string was replaced by the string "NIL". Queries with no answer were not eliminated: on the contrary, twenty NIL questions were included in the final version of each monolingual test set to evaluate systems' accuracy in recognising that there was no response. In the English version of each question, a default negative value "-1" was assigned to distinguish it from the zero used in NIL questions.

- the answer string represents the correct and exact answer found by the assessors, and supported by the document returned in the attribute `idx`. The string “SEARCH” followed by the translation within square brackets of the correct answer found in the source corpus constituted a valuable help for the assessors who would process the shared questions. The merging revealed that 246 questions had at least one answer in all the three reference document collections, 111 had at least a response in two of them, and the remaining 93 just in the source corpus in which they were first processed.

Example of a question in the DISEQuA Corpus:

```
<qa cnt="24" type="MEASURE">
  <language val="ITA" original="TRUE">
    <question assessor="Ale-irst">
      Quanti abitanti ha l'Iraq?
    </question>
    <answer n="1" idx="SDA19940225.00018">
      11 milioni
    </answer>
  </language>
  <language val="SPA" original="FALSE">
    <question assessor="Anselmo-UNED">
      ¿Cuántos habitantes hay en Irak?
    </question>
    <answer n="1" idx="EFE19941110-06054">
      18 millones
    </answer>
  </language>
  <language val="DUT" original="FALSE">
    <question assessor="LIT-UVA">
      Hoeveel inwoners heeft Irak?
    </question>
    <answer n="1" idx="AD19941012-0075">
      ongeveer 20 miljoen
    </answer>
  </language>
  <language val="ENG" original="FALSE">
    <question assessor="">
      How many inhabitants does Iraq have?
    </question>
    <answer n="1" idx="-1">
      SEARCH [11,000,000]
    </answer>
  </language>
</qa>
```

For further information about the DISEQuA corpus, and in particular about the procedure used during the verification of the queries, see the paper “Creating the DISEQuA Corpus: a Test Set for Multilingual Question Answering”, in the Working Notes of the CLEF 2003 Workshop. (URL: <http://clef.iei.pi.cnr.it:2002/>)