Textual information retrieval systems test : the point of view of an organizer and corpuses provider.

Patrick Kremer & Laurent Schmitt

INIST / CNRS
2, allée de Brabois
54514 Vandoeuvre-lès- Nancy
patrick.kremer@inist.fr, laurent.schmitt@inist.fr, amaryllis@inist.fr

Abstract

Amaryllis is an evaluation programme for text retrieval systems which has been carried out as two test campaigns. The second Amaryllis campaign took place in 1998/1999.

Corpuses of documents, topics, and the corresponding responses were first sent to each of the participating teams for system learning purposes. Corpuses of new documents and a set of new topics were then supplied for evaluation purposes. Two optional tracks were added for Internet and interlingual track.

The first track of these contained a test via the Internet. INIST sent topics to the system and collected responses directly, thus reducing the need for conceptor manipulations.

The second contained tests in different European Community language pairs. The corpuses of documents consisted of records of questions and answers from the European Commission, in parallel official language versions. Participants could use any language pair for their tests.

The aim of this paper is to give the point of view of an organizer and corpus provider (INIST) on the organization of an operation of this sort. In particular, it will describe the difficulties encountered during the tests (corpus construction, translation of topics and systems evaluation), and will suggest avenues to explore for future tests.

1. Introduction

Amaryllis is an evaluation programme for text retrieval systems . This project was co-sponsored by AUF (formerly Aupelf-Uref) and the French Ministry of Education, Research and Technology. The program was carried out as two test campaigns. The first, which took place in 1996/1997, was an exploratory learning campaign, while the second, in 1998/1999 was the test phase proper.

The exploratory phase aimed to establish close links with the developers of information retrieval systems, to build up an initial document corpus in French, and to "test" the systems using existing evaluation methodology (TREC : Harman, 1994). This phase was therefore considered as a "modelling" cycle. It was concluded at the JST in Avignon (Coret & al., 1997).

For these tests, corpuses of documents (articles from "Le Monde" and titles and abstracts of scientific articles from INIST databases), topics, and the corresponding responses were given to each of the participating teams for system learning purposes. Corpuses of new documents and a set of new topics, without the corresponding responses, were then supplied for evaluation purposes, i.e. tests simulating a routing task using old topics with new documents, and other tests simulating a search (an ad hoc TREC task) using new topics with old documents. Two graphs were plotted on the basis of the TREC evaluation program to show degrees of precision with respect to the number of documents retrieved, and pecision with respect to recall ability. The main difference between Amaryllis and TREC was the comparison of participants' responses with "correct" responses supplied by the document providers (OFIL[1] and INIST[2]) using their own methods and tools.

The second phase (1998/1999) was similar to the first, but used new corpuses of documents, topics and responses (supplied by the corpuses providers). The documents were again items from "Le Monde" and titles and abstracts of scientific articles from INIST databases, but also included monographs on Micro-Melanesia provided by the LRSA[3].

Two optional tracks were also added for this second test campaign: an Internet test track and a interlingual track.

The first was designed to test topics with documents through the Net. The organiser (INIST) sent topics to each system via Internet and collected the responses directly. This may be viewed as an attempt to establish a basis for testing from an end-user point of view.

The other track was designed to test systems for different pairs of European Community languages , mainly English , French, German, Italian, Portuguese and Spanish. The documents were records of questions and answers from the European Commission in parallel official language versions (ELRA corpus). Participants were free to use any language pair (i.e. English topics with French documents or Portuguese topics with German documents). Their responses were compared to the "correct responses".

The aim of this paper is to give the point of view of an organiser and corpus provider (INIST) on

[1] Observatoire Français et international des Industries de la Langue

[2] INstitut de l'Information Scientifique et Technique

[3] LRSA : Laboratoire de Recherches Sémiographiques en Anthropologie – Université de Laval , Canada

the organisation of an operation of this type. In particular, it will highlight the difficulties (or advantages) that emerge in carrying out evaluation tests of this kind or in building up the tools to be used in the evaluation tests : construction of the corpuses of material (documents, topics and correct responses), translations for interlingual tests, evaluation methods and Internet tests. It will also highlight questions raised by the tests and suggest guidelines and ideas for future testing.

2. Protocol

Like the first phase of the Amaryllis experiment, the second closely followed TREC procedures. The differences were essentially as follows:

- corpus volumes were substantially lower than in TREC (several Gbytes for the documents and several hundred topics).
- systems were tested separately with each corpus
- "correct" response references were supplied before the tests. In TREC, the references were obtained by compiling responses supplied by each system and subsequently selected by an evaluation team, whereas in Amaryllis, an evaluation team built up a “correct” response corpus before the test. This corpus was re-evaluated after the test against responses from participants

2.1. First phase: system learning

The participants received the Amaryllis Volume 3 CD-ROM containing:

The OFIL Corpus (newspaper articles from “Le Monde”)
OD1: 10,500 articles (approx.), 33.4 MB,
OT1: 26 topics
OT1D1: "Correct" responses from the supplier
The INIST Corpus (scientific titles and abstracts from Inist databases)
ID1: 151,000 notices (approx.), 64.5 MB
IT1: 30 topics
IT1D1: "Correct" responses from the supplier
LRSA Corpus (Monographs on Micro Melanesia)
EM1: 6 monographs , 2.9 MB
MT1: 11 topics
MT1D1: "Correct" responses from the supplier
The ELDA Corpus (Parallel interlingual document corpus from the European Commission)
ED1: 3.511 entries x 6 languages (approx.), 54.5
ET1: 15 topics x 6 languages
ET1D1: "Correct" responses from the supplier

The participating teams optimised their systems and formulated requests from the OT1 and IT1 search topics (to be used for the classification tests in the second phase). A time-scale of 3 months was planned for this initial phase.

There were two possible methods for formulating requests from the topics:

- in automatic mode where the system creates the request automatically from one or more search topic items,
- in manual mode where the request is constructed by the user alone or with system assistance

At the end of this learning period, each testing team sent the organiser the first 250 sorted responses.

No modifications of the systems were allowed after this learning period and until the evaluation results were supplied.

2.2. Second phase : evaluation Tests

Two kinds of tests were used : classical tests carried out by the participant with new corpuses of documents or topics, and optional tests carried out by the organizer (INIST) through the net before any test??.

For this phase, the participants received the CD-ROM containing Amaryllis Volume 4.
Volume 4 contained:
The OFIL Corpus
OD2: 9,300 articles (approx.), 30 MB,
OT2: 26 topics
The INIST Corpus
ID2: 130,600 entries (approx.), 64.5 MB
IT2: 30 topics
The LRSA Corpus
EM2: 6 monographs , 1.2 MB
MT2: 10 topics
The ELDA Corpus
ED1: 3.511 entries x 6 languages (approx.), 54.5
ET2: 15 topics x 6 languages

2.2.1. Classical tests

Two types of evaluation test were carried out:

- routing tests: the formulations for the OT1 and IT1 topics (sent to the organizer at the end of the first phase) were "applied" with no modification to the new document corpuses, OD2 and ID2 respectively,
- tests simulating a search: the new topics, OT2 and IT2, were "applied" to the old documents, OD1 and ID1 respectively.

For these search tests, the participants could choose to formulate search requests either automatically or manually as in the classification tests, and also with feedback (the initial request being formulated either manually or automatically: analysis of the relevant documents found being used to refine the request either manually or automatically).

At the end of each testing phase, each participant sent the organizer:

Files containing the first 250 responses sorted by relevance,

A completed questionnaire, intended to provide a deeper understanding of the functioning of each system and the work carried out.

2.2.2. Interlingual tests

For the interlingual test, only a search task was carried out, using the same documents as in the learning phase but with new topics.

2.2.3. Internet task

For the internet test, participants had to create databases with the documents, then incorporate their API (Application Programming Interface) into a CGI (Common Gateway Interface), after which the test was conducted through the Net by INIST. These tests simulated an automatic search.

2.3. Analysis of Results

The results files from each participant (all topics) were processed with TrecEval software to produce (amongst other output) graphs of two types :
- precision as a function of the number of documents found (5 first, 10 first, ... , 1,000 first).
- precision as a function of recall.

These two graphs were first plotted using the initial references provided by the suppliers, and subsequently revised as a function of the participants' responses to produce a set of reference responses.

2.4. Construction of the Corpuses

2.4.1. Documents

For all tests except interlingual tests, the text documents were in French.

The documents were:
- articles (titles and texts) from "Le Monde", with each corpus of documents supplied by OFIL covering a three-month period (01-01-93 / 31-03-93, 01-04-93 / 30-06-93),
- titles and summaries of scientific articles covering all subjects, extracted from the Pascal (1984 to 1995) and Francis (1992 to 1995) bibliographic databases supplied by INIST.
- Six monographs on Micro-Melanesia from LRSA.

The corpus used for the interlingual tests was extracted from the ELRA catalogue of Interlingual Corpora for Co-operation, in the 14 languages of the European Community. This corpus was held by INIST for future processing purposes. Six languages were used in the test : English, Spanish, French, German, Italian and Portuguese.

All the document corpuses were in SGML format with iso-latin coding.

The documents were structured according to a simplified DTD supplied by TEI[4], which included the management of the logical structure of a book. This enabled the corpus of books from the LSRA to be included.

2.4.2. Search Topics

These were derived from real requests made by end users, and in principle included all the informational elements required to understand the fields covered and to evaluate their relevance.

They included the following information:

Field: to define the knowledge field to which the topic belongs
Subject: a title defining the topic
Question: i.e., the user's request
Additional information: to provide specific information on which documents in the corpus should be kept
Concepts: containing a group of key words to limit the search area.

Example of an INIST Topic

```
<record>
<num>15</num>
<dom>Médecine</dom>
<suj>Ulcères gastroduodénaux</suj>
<que>Traitements chirurgical ou médicamenteux des ulcères gastriques et duodénaux</que>
<cinf>Pour être pertinent, un document décrira une technique, un résultat d'un traitement soit médicamenteux soit chirurgical ou les 2 associés, d'un ulcère dont la localisation pourra être gastrique et/ou duodénale</cinf>[5]
<ccept>
<c>Chirurgie</c>
<c>Chimothérapie</c>
<c>Antiulcéreux</c>
</ccept>
</record>
```

Topics were built up by:
- OFIL, with the help of documentalists from Le Monde, using requests made by journalists,
- documentalists from INIST (specialising in the relevant fields) using requests made by their end-users,
- specialists on Micro-Melanesia for the LRSA,
- documentalists from INIST for the interlingual corpus.

These topics had to cover different fields and produce a significant number of relevant responses from each document corpus. They were tested to ensure that they produced enough responses. This led to some of them being modified or added to.

2.4.3. Corpus of reference responses

The response files contained a list of the numbers of all the relevant documents for each topic, the latter being identified by a number.

We wanted to establish a corpus of reference responses before the participants began testing .

The responses were supplied by the providers (OFIL, LRSA and INIST) using their own methods and tools (initial reference):

At OFIL, with the help of documentalists from the newspaper Le Monde

At INIST, where document engineers specialising in the relevant field made a pre-selection, deliberately making it as wide as possible, using the titles and abstracts together with the keywords and classification codes which appear in the Pascal and Francis data bases. The keywords and classification codes were not given to the participants. The list of documents thus obtained was then sorted manually by the engineers, with extra weighting given to those that most exactly answered the question asked.

Two types of reference were constructed in this way for INIST as an example : IT1D1 (T1 topics with the D1 documents) for the learning phase, with no subsequent modification allowed, and IT2D1 (T2 topics with D1 documents) and IT1D2 (T1 topics with D2 documents) for the test phase.

The IT1D2 and IT2D1 references were modified manually by INIST in the light of participants' responses. This allowed us to improve the quality of the reference responses. We did this because the relevance of the results is not absolute : the first filtering of the documents may be flawed, thus omitting possible responses.

[4] TEI: Directives for SGML encoding of text to facilitate exchanges and automatic processing.

[5] To be considered relevant, a document must describe a technique or the results of a medical and/or surgical treatment of a gastric or duodenal ulcer, or both.

This adjustment process took place in three stages:
- review of the documents found by more than half of the participants, but not by the supplier.
- review of the documents contained in the references but not found by any participant.
- review of the first ten documents found by each participant but not by the provider, to take into account the ranking of responses rather than the number of times they were found by the different systems.

This resulted in modifications in the form of addition and removal of documents from the corpus of reference responses. This method remains open to question, as indeed must all attempts to create a reference.

Rather than making a single reference a posteriori using the participants' responses , a reference was made a priori and then modified a posteriori so as to obtain an optimum number of relevant documents that could be located by any individual system.

3. INIST experience

In each task (classical, interlingual or Internet test), difficulties were encountered in the organisation or supply of corpuses. We will now review these as encountered in the different tests.

3. 1. Classical tests

Finding a corpus of documents on the information market is not a problem : today, everyone has them. Many newspapers, databases and other organizations could export their archives: any corpus is relatively easy to obtain when it is tagged in SGML format .

Finding a corpus of documents should not be a problem, as owners have had to invest their time in formatting data for their archives.

But for an organizer of an operation such as Amaryllis it is very difficult to obtain queries and responses. Corpus providers do not have queries or do not keep archives of queries (topics), and still less the responses to these queries. When they are needed, creating them (or testing them when they can be found) involves a considerable amount of time and effort. Queries have to be created, selected and tested to retrieve enough responses and to formulate topics with all the different fields described above., Providers of queries and documents then have to build up corpuses of reference responses, which takes a long time as numerous adjustments have to be made. In 1998, when we tried to find new material for the second phase, we were unable to find any provider in France that could supply documents, queries and responses.

In fact, the advantage of this method of providing reference responses is that corpuses of correct response are built up according to the contents and not only on the basis of the responses supplied by the participants' systems. However, it increases the work involved, and only can be applied to small corpuses. When their size increases, and when they are very large we recommend using a different method to generate responses, like the TREC method or a pooling method with tags, as used in the GRACE project (Adda & al., 1998), where 1 million words were tagged out of a total of 10 million.

As one of the organizers INIST experienced considerable difficulty in obtaining document corpuses and the manpower required to generate added value by creating topics and reference responses. When this was not possible, we were at least able to act as a provider of topics and responses within our field of competence.

3.2. Interlingual tests

This test had two objectives: to compare the performances of a retrieval system in different languages, which was more a way of testing linguistic resources added to the system, and to test its translating capacities when retrieving documents in one language in response to a query in another language.

The difficulties encountered in the interlingual task related not only to the test corpus, but also to translation.

Like in the classical tests, it was relatively easy to find different corpus in different languages (newspapers from different countries), but we also had to find queries and responses to those queries. This meant setting up an organisation in different countries with a network of correspondents, which was not an easy task.

We had the possibility of acquiring aligned texts from ELDA in different European Community languages . Such documents corpuses are invaluable though not readily found, and we were unable to obtain the necessary queries and responses as well.

We therefore had to create queries ourselves in order to test them, so that INIST effectively became the provider of both topics and responses for the interlingual test. This was possible because the documentalists at INIST were able to process topics from the European Commission.

In our case, the queries were made in French for French documents. As we received the responses to queries, we translated the topics into the other 5 languages to obtain a single reference corpus to limit the variations due to translators.

To avoid problems stemming from the translation of topics, we used native speakers and gave them responses in the different languages to help them to harmonise the vocabulary by using the same words in the topics as in the ELDA documents .

3.3. Internet tests

There were no problems with the corpus of material because we used the same corpuses as in the classical tests.

The objectives were not to test the possibilities of the system through the net but rather to test the system with no manipulation by the system owner. Testing a system through the net is only a technical matter (incorporation of an API into a CGI) but this kind of test can be seen as a system test from the end-user point of view. When testing a system using classical methods, we are testing not only the intrinsic value of the system, but also the linguistic resources added to it and the added value of the conceptor.

The intrinsic value of the system is represented by the search engine itself, by index constitution and by the linguistic model use. The ressources values are the grammars and dictionaries added. The conceptor value is the reformulation of the querries, the ponderation used, the time past by the conceptor and by the system to work and learn and the system knowledge by the conceptor.

The first two items (intrinsic value of the system and linguistic resources) represent the “actual” performance of the system, but the third represents an individual contribution. This kind of test effectively discounts the conceptor’s contributions (time spent on searches, query reformulation and system learning), thus placing all systems in an identical situation. To make further progress, what is needed is a search group of people from different professional background who would test different systems at the same time.

Such a group of end-users would not be easy to get together, while selecting profiles and comparing systems against evaluation procedures would be a major project , which needs to be addressed with other research laboratories in the world.

4. Conclusion

Tests like Amaryllis (or TREC) are not easy to build. Document corpuses have to be located and others built up. To make progress, INIST needs to work in collaboration with others: we need new corpuses of material and methodological refinements in order to explore new avenues or to make further progress in the directions we are already working in (panel of end-users and translations).

5. References

Adda G., Mariani J., Lecomte J., Paroubek P., and Rajman M. (1998). The GRACE French part-of-speech tagging evaluation task. In A. Rubio, N. Gallardo, R. Castro, and A. Tejada, editors, *First International Conference on Language Resources and Evaluation*, volume I, pages 433-441, Granada.

Coret A., Kremer P., Landi B., Schibler D., Schmitt L., Viscogliosi N. (1997). Accès à l’information textuelle en français : le cycle exploratoire Amaryllis . In 1re JST Francil, 5-8.

Harman D. (1994). Overview of the second Text Retrieval Conference (TREC-2). In Harman DK, ed. Proceedings of the second Text Retrieval Conference. NIST Special Publication 500-215, 1-20.