Arabic, English and French: three languages in a filtering systems evaluation project

Romaric Besançon[1], Djamel Mostefa[2], Ismaïl Timimi[3], Stéphane Chaudiron[3], Mariama Laïb[1] and Khalid Choukri[2]

[1]CEA-LIST
18, rue du panorama
BP6 92265 Fontenay aux Roses, France
{romaric.besancon,meriama.laib}@cea.fr

[2]ELDA
55-57 rue Brillat Savarin Paris, France
{choukri,mostefa}@elda.org

[3]Université de Lille3 – GERiiCO
BP 60149- 59653
Villeneuve d'Ascq cedex France
{stephane.chaudiron,ismail.timimi}@univ-lille3.fr

Abstract

The InFile project (INformation, FILtering, Evaluation) is a cross-language adaptive filtering evaluation campaign, sponsored by the French National Research Agency. The project is organized by the CEA-LIST, ELDA and the Laboratory GERIICO of the University Lille3. It has an international scope as it was a pilot track of the CLEF 2008 and a main track of the CLEF 2009 campaigns. The corpus is a collection of about 1,4 millions newswires in three languages, Arabic, English and French provided by Agence France Press (AFP) and selected from a 3 years period. The profiles' corpus (the corpus requests) is made of 51 profiles from which 30 concern general news and events (national and international affairs, politics, sports…) and 21 concern scientific and technical information. This paper is presenting the InFile evaluation paradigm in general and focuses on a study of the Arabic part of the corpus in particular. The coverage mismatch between profiles and Arabic documents, conceptual and terminology gaps in the transfer between English/French and Arabic are also discussed in this article.

Introduction

The InFile[1] project (INformation, FILtering, Evaluation) is organizing for the second time a cross-language adaptive filtering evaluation campaign. The project is sponsored by the French National Research Agency and is organized by the CEA LIST, ELDA and the University Lille3 (Lab. GERiiCO). The campaign has an international scope as it is a main track of the next CLEF 2009 campaigns and benefits from the NIST TREC endorsement to promote InFile especially towards American participants during the 2008 TREC workshop, last November.

Information filtering systems may be used in different business contexts of use : for example, text routing which involves sending relevant incoming data to individuals or specific groups, categorization process which aims at attaching one or more predefined categories to incoming documents, or anti-spamming which tries to remove « junk » e-mails from the incoming e-mails.

For the InFile project, we retained the context of competitive intelligence in which the information filtering is a very specific subtask of the information management process (Bouthillier, 2003). In this approach, the information filtering task is very similar to Selective Dissemination of Information (SDI), one of the original and usual function assumed by documentalists and, more recently, by other information intermediaries such as technological watchers or business intelligence professionals.

In this communication, we present the main characteristics of the next InFile campaign which will be held in 2009 according to the CLEF calendar and an analysis of the specificity of dealing with the Arabic language. The coverage mismatch between profiles and Arabic documents, conceptual and terminology gaps in the transfer between English/French and Arabic are discussed in the second part of the article.

InFile evaluation campaigns

We summarize here the main characteristics of the campaigns.
Detailed information is available in previous publications (Besançon & al, 2008), (Chaudiron & al, 2008).

Goals of the campaign

The InFile evaluation campaign (Besançon, 2008) measures the ability of filtering systems to successfully separate relevant and non-relevant documents in an incoming stream of textual information. Following Belkin and Croft (Belkin, 1992), an information filtering system

[1] http://www.infile.org

is a system designed to manage unstructured or semistructured data. Information filtering systems deal primarily with textual information, involving large amounts of data incoming through permanent streams such as newswire services. Filtering is based on individual or group information profiles which assume to represent consistent and long-term information needs. From the user point of view, the filtering process is usually meant to extract relevant data from the data streams, according to the user profiles.

This means that we are paying a particular attention to the context of use of filtering systems by real professional users. Even if the campaign is mainly a technological oriented evaluation process, we have adapted the protocol and the metrics, as close as possible, to what we call the « ground truth » or « real uses ».

So, the goal of the InFile project is twofold: first and mainly, it is evaluation campaign involving academic and industrial participants; secondly, it is an attempt to better understand and model the human information filtering practices and possibly « to translate » it in evaluation protocol and metrics.

Previous evaluation campaigns have been proposed in the past years on Adaptive Filtering systems, including the Text Retrieval conference (TREC) Adaptive Filtering tracks from 2000 to 2002 (Roberston, 2002) and the Topic Detection and Tracking (TDT) campaigns from 1998 to 2004 (Fiscus, 2004). The InFile campaign can mainly be seen as a cross-lingual pursuit of the TREC11 Adaptive Filtering task, with a particular interest in the correspondence of the protocol with the ground truth of competitive intelligence professionals. In the TDT campaigns, focus was mainly on topics defined as "events", with a fine granularity level, and often temporally restricted, whereas in InFile (similar to TREC11), topics are of long term interest and supposed to be stable, which can induce different techniques, even if some studies show that some models can be efficiently trained to have good performance on both tasks (Yang 2005).

Main characteristics of the campaign

The InFile evaluation campaign is:

- crosslingual : Arabic, English and French are concerned by the process but participants may be evaluated on mono, bilingual or mutlilingual runs ;
- the corpus is composed of around 100,000 newswires selected from the Agence France Presse (AFP) stream. There are two groups of topics, one concerning general news and events, and a second on Scientific and Technical Information ;
- the evaluation task is performed either (1) using an automatic interrogation of participating systems with a simulated user feedback (system are allowed to use the feedback at any time to increase performance) or (2) in batch mode for which all documents are released at the same time ;
- for each task (interactive filtering versus batch mode filtering) systems have to provide a Boolean decision for each document according to each profile.

Description of the protocol

Before the evaluation run, some general information about the two domains of interest are given to the participants in order to adapt their systems, if necessary.
Moreover a development set of two profiles with relevant documents has been made available few months before the evaluation starts.
When the evaluation starts, profiles are given to participants. Each profile is composed of 5 information fields : a short title, a descriptive sentence of the subject, a longer description of what is (or not) a relevant document, a set of keywords (5 max.), and a sample excerpt of relevant document for the topic (found on the web and not taken from the news collection to be filtered).

Interactive filtering evaluation protocol
The goal of the interactive filtering track is to test crosslingual adaptive filtering systems in a way as close as possible to the usage of such systems by competitive intelligence practitioners. In particular, the collection of documents to filter is not made available in a single archive: the documents are made available one at a time, in an interactive protocol. A client-server architecture has been developed to support this protocol:
A document server is deployed at ELDA : documents are retrieved from this server, and the results of the filtering are sent back to the server. Participant systems communicate with this server using a web service protocol. The evaluation works as follows:

1. the participant filtering system registers to the document server
2. it retrieves a document
3. it compares the document to each topic, and for each topic for which the document is relevant, it sends the result (pair document-topic) to the server
4. for each result sent, the filtering system can ask the server for feedback : the server returns a Boolean answer indicating if the document is indeed relevant to the topic (according to *a priori* assessments). Feedback is given only for the results sent, i.e. participants cannot have feedback on a negative result (i.e. participant cannot ask “am I right to discard this document for this topic ?”). There is a limited number of feedbacks allowed.
5. a new document can be retrieved (back on 2.)

On the server side, results are gathered in a run file, that will afterwards be evaluated according to *a priori* assessments, and using several evaluation measures, such as precision, recall, F-measure, utility measure, detection cost etc... *A posteriori* assessments will be performed for

documents found relevant for most participant systems and not previously assessed.

Batch filtering evaluation protocol
For the batch mode filtering track, all documents are sent to the system and for each document and each profile, a Boolean decision must be returned by the system.
The news collections are given to the participants in three archives, one for each language and the 50 topics are also distributed in three XML files, one per language.
Then the participant must compare each topic in a source language to the documents in the target languages. Every source/target languages are allowed (monolingual filtering, crosslingual filtering or multilingual filtering).

Metrics

The results returned by the participants are binary decisions on the association of a document with a profile. The results, for a given profile, can then be summarized in a contingency table of the form:

	Relevant	Not Relevant
Retrieved	a	b
Not Retrieved	c	d

Table 1Contengency table

On these data, a set of standard evaluation measures is computed: *Precision* (P=a/a+b), *Recall* (R=a/a+c), and *F-measure* (F=(1+α)PR/αP+R) (Van Rijsbergen, 1979).

Following the TREC Filtering tracks (Hull,1999) (Robertson,2002) and the TDT 2004 Adaptive tracking task (Fiscus, 2004), we also consider the *linear utility*, defined as $u = w_1 \times a - w_2 \times b$, where w_1 is the importance given to a relevant document retrieved and w_2 is the cost of an irrelevant document retrieved, and its normalized version:

$$u_n = \frac{\max(u/u_{\max}, u_{\min}) - u_{\min}}{1 - u_{\min}}$$

where u_{max} is the maximum value of the utility and u_{min} a parameter considered to be the minimum utility value under which a user would not even consider the following documents for the profile.

For sake of comparison, we will also compute the *detection cost*, which is the standard measure from the Topic Detection and Tracking campaigns (TDT2, 1998), and depends on given costs for missed documents and false alarms.

The average scores on all profiles are the macro-averaged values of the considered scores, which gives equal weight to each profile.

In order to measure the adaptivity of the systems, the measures are also computed at different times in the process (e.g. every 10 000 documents), and an evolution curve of the different values across time is proposed. Additionally, two following experimental measures are tested: the first one is an *originality* measure, defined as a comparative measure corresponding to the number of relevant documents the system uniquely retrieves (among participants). It gives more importance to systems that use innovative and promising technologies that retrieve "difficult" documents. The second one is an *anticipation* measure, motivated in competitive intelligence by the interest of being at the cutting edge of a domain: it is measured by the inverse rank of the first relevant document detected (in the list of the documents), averaged on all profiles. The measure is similar to the mean reciprocal rank (MRR) used for instance in Question Answering Evaluation (Voorhees, 1999), but is not computed on the ranked list of retrieved documents but on the chronological list of the relevant documents.

The corpus

In order to ensure that none of the participants already worked on the test corpus in previous evaluation campaigns, the InFile project has created a very new test corpus.

This corpus is composed of a collection of 100,000 recent newswires of general and scientific interest from Agence France Presse. This collection is composed of three sets:

- a set of relevant documents provided by information professionals (assessors) ;
- a set of non relevant but close-to-profile documents specially chosen to ensure some confusion with the profiles ;
- and the rest of the collection which is a large set of irrelevant documents in which the two previous subsets are hidden. The size of this subset is large enough to prevent any manual examination of the corpus. In order to ensure that this corpus does not contain any relevant document, assessors use a state-of-the-art information retrieval system on the full data to detect and eliminate such documents.

The profiles (requests) and the repository (pertinent documents) have been created by information professionals and practitioners.

The approach for building the test collections is depicted by Figure 1.

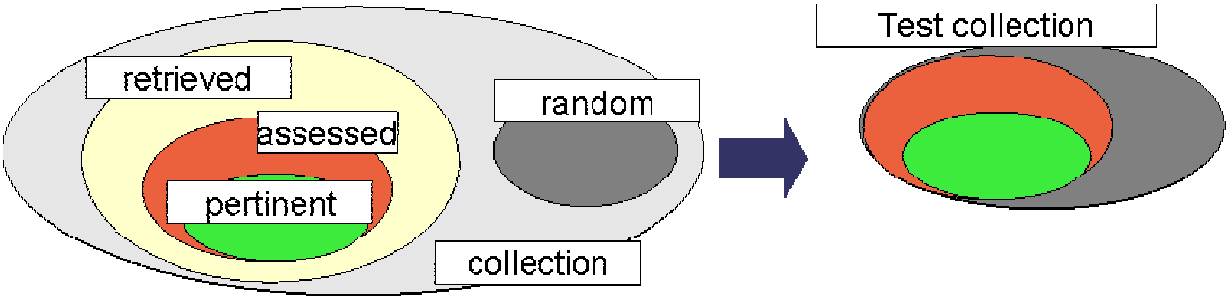

Figure 1 Test data collection construction

Difficulties and specificities of the Arabic language

The Arabic part of the corpus is made of news wires from AFP and profiles used for the evaluation.
Profiles were produced by information professionals and practitioners.
For practical reasons, profiles were created in French or in English directly and then translated into Arabic. Only the <sample excerpt> field was not translated from English or French to Arabic. We recommended that the translation from French to English (resp. English to French if the profile was first written in English) was produced by the practitioner or information professional who created the profile in French (resp. English). This enables a maximum preservation of sense and respects the terminology across different domains and languages.

For evident reasons, the translation to Arabic could not be done directly by the same professionals who wrote the profiles in French and in English. We recruited Arabic translators with sufficient backgrounds in business intelligence and multilingual information retrieval to carry out this work. Since this task is quite important for the sake of the evaluation, we selected Arabic translators with a good knowledge of both English and French. However, given the fact that no Arabic terminology lexicon covering all scientific and technical terms was available, we had to face some difficulties for the translation of certain concepts and terms, especially for the emerging ones. So we asked several translators of universities and professionals working in the Arabic region with different variants of Arabic (Egypt, Lebanon and Morocco) to carry out this task; we selected Egypt, Lebanon and Morocco as Arabic, French and English are widely spoken in these countries.
To deal with less known concepts or non official terms, some translators had to use various tools and sources such as non specialized lexicons, study of the frequencies of terms on the web, comparison of available translations within the free encyclopedia Wikipedia, etc. To reduce the impact of terminology differences across various sources, we decided to keep at the end in the profile's description the most common term on the web and to write between brackets the other synonym terms used in other sources or contexts.

For clarity, we present here some examples of difficulties we had to face when translating to Arabic:

- acronyms that are not used in Arabic: DGN, CIO
- terms that are sometimes translated and sometimes transcribed (تكنولوجيات – تقنيات)
- concepts that are used very locally : part-time work
- emerging concepts: m-commerce (الجوّالة- الخلوية - التجارة الإلكترونية)
- unknown concepts: cosmetofoods

Conclusion

In this paper we presented the InFile project, a cross lingual filtering evaluation campaign for Arabic, English and French. The evaluation protocol and ground truth were created with the idea to be as close as possible to the real usage. We also studied the Arabic part of the corpus and explained the gaps and problem that may arise when trying to transfer information from English and French to Arabic.
Beyond these issues that may arise from the Arabic language in such evaluation projects of information technology, this campaign has enabled us to point out some limitations of Arabic language resources and the need to rethink the construction of reliable and well known resources, especially in the field of information science and technology, where technological innovation and therefore terminology innovation is perpetual.

All material produced for the InFile campaigns will be made available as an evaluation package through ELRA's catalog (http://catalog.elra.info). The evaluation package will consist of the AFP corpus used during the evaluation, the profiles, the assessments, the protocols and tools developed for the campaign. The evaluation package will enable an external player to evaluate its technology offline and compare its results with those of the participants.

Acknowledgements

The InFile project is partially funded by the Agence Nationale de la Recherche (ANR-06-MDCA-011-01). We also thank all the information professionals from INIST,

ARIST, Digiport and ONERA who constructed the profiles.

Bibliographical References

Belkin N., Croft B. (1992). Information filtering and information retrieval : two sides of the same coin. In *Communications of the ACM*, december 1992, vol. 35, n°12, p. 29-38.

Besançon R., Chaudiron S., Mostefa D., Hamon O., Timimi I., Choukri K.(2008). Overview of the CLEF 2008 INFILE Pilot Track. In *Working Notes of the Cross Language Evaluation Forum (CLEF 2008)*, Aarhus, Sept. 2008.

Besançon R., Chaudiron S., Mostefa D., Timimi I. Choukri K. (2008). The InFile project: a crosslingual filtering systems evaluation campaign. In *Proceedings of LREC 2008*, Marrakech, Morocco, May 2008.

Bouthillier F., Shearer K. (2003). *Assessing Competitive Intelligence Software : A Guide to Evaluating CI Technology*, Medford, Information Today Inc., 2003.

Chaudiron S., Besançon R., Mostefa D., Timimi I. Laib, M., Choukri K. (2008). Le filtrage automatique de l'information multilingue : une évaluation inspirée des vérités terrain. In Actes du *Colloque International en traductologie et TAL*, Oran, Algeria, June 2008.

Fiscus, J.G, Wheatley, B (2004) Overview of the TDT 2004 evaluation and results, In TDT'02, NIST.

Hull D., Roberston S. (1999) The TREC-8 Filtering Track Final Report, in *Proceedings of the Eighth Text REtrieval Conference* (TREC-8)

Robertson S., Soboroff I. (2002). The TREC 2002 Filtering Track Report. In *Proceedings of The Eleventh Text Retrieval Conference (TREC 2002)*. NIST Special Publication : 500-251

TDT2 (1998) The Topic Detection and Tracking Phase 2 (TDT2) Evaluation Plan, NIST http://www.nist.gov/speech/tests/tdt/1998/doc/tdt2.eval.plan.98.v3.7.pdf

Van Rijsbergen, C.J. (1979) *Information Retrieval*. Butterworths, London.

Yang Y., Yoo S., Zhang J., Kisiel B. (2005) Robustness of adaptive filtering methods in a cross-benchmark evaluation, In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, Salvador, Brazil, pp. 98 - 105