Dealing with MultiLingual Information Access: Grid Experiments at TrebleCLEF

Nicola Ferro* and Donna Harman†

*Department of Information Engineering – University of Padua – Italy
ferro@dei.unipd.it

†National Institute of Standards and Technology – USA
donna.harman@nist.gov

***Abstract*— This paper discusses some issues related to the planning of a series of *grid experiments* in the context of the TrebleCLEF project with the aim of improving the comprehension of *MultiLingual Information Access (MLIA)* with respect to languages, e.g. differences in the effectiveness of the various weighting schemes and retrieval models.**

I. Introduction

During the workshop on "The Future of Large-scale Evaluation Campaigns"[1] [1], which was organised jointly by the University of Padua and the DELOS Network of Excellence on Digital Libraries[2] and held in Padua, Italy, March 2007, a critical assessment of the scientific results of the *Cross-Language Evaluation Forum (CLEF)* initiative[3] has been conducted and recommendations for TrebleCLEF[4], a coordination and support action funded by the European Community within the Seventh Framework Programme, have been provided.

TrebleCLEF intends to promote research, development, implementation and industrial take-up of multilingual, multimodal information access functionality in the following ways [2]:

- by continuing to support the annual CLEF system evaluation campaigns with tracks and tasks designed to stimulate R&D to meet the requirements of the user and application communities;
- by constituting a scientific forum for the MLIA community of researchers enabling them to meet and discuss results, emerging trends, new directions;
- by acting as a virtual centre of competence providing a central reference point for anyone interested in studying or implementing MLIA functionality and encouraging the dissemination of information

In this context, the possibility of running a series of *grid experiments* to improve the comprehension of *Information Retrieval (IR)* and *MultiLingual Information Access (MLIA)* with respect to languages has been discussed. Indeed, individual researchers or small groups do not usually have the possibility of running large-scale and systematic experiments over a large set of experimental collections and resources in order to improve the comprehension of MLIA systems and gain an exhaustive picture of their behaviour with respect to languages. Therefore, a series of systematic grid experiments can re-use and exploit the valuable resources and experimental collections made available by CLEF in order to gain more insights about the effectiveness of the various weighting schemes and retrieval techniques with respect to the languages and to disseminate this knowledge to the relevant application communities, such as the *Digital Library (DL)* one.

This kind of experiments has been proposed with some goals in mind:

- to look at differences across a whole set of languages;
- to come out with best practices for each language;
- to help other countries to bring up their expertise in the IR field and create IR groups;
- to serve as a resource place, where all this information is managed and made available by means of some proper information management system.

Moreover, these experiments could be conducted by building on the methodology adopted by [3], who compared different stop lists, stemmers, weighting schemes, query translation and merging strategies across five European languages (English, French, Italian, German, and Spanish).

The paper describes the methodology and the experimental set-up according to which we plan to carry out these experiments. The paper is organized as follows: Section II introduces the proposed model for grid experiments; Section III presents some possibilities for visualizing the experimental results; Section IV discusses how grid experiments can promote technology transfer; finally, Section V wraps up the discussion and outlines the future work.

II. Modelling Grid Experiments

As discussed in Section I, the overall goal is to analyse how retrieval effectiveness is affected by the interaction of the different components of an *Information Retrieval System (IRS)* with respect to the language at hand[5].

[1] http://ims.dei.unipd.it/events/2007/future-evaluation-campaigns/future-eval-index.html
[2] http://www.delos.info/
[3] http://www.clef-campaign.org/
[4] http://www.trebleclef.eu/

[5] Note that understanding the impact of language changes over the time - e.g. from Shakespeare English to today English - is out-of-scope for this study.

We plan to face this problem in the context of laboratory experimentation, according to the Cranfield evaluation methodology [4] which makes use of experimental collections consisting of documents, topics, and relevance judgements. In particular, the CLEF collections will be exploited to conduct these experiments in the context of a traditional ad-hoc task.

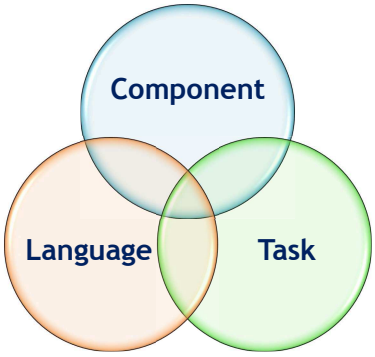

Fig. 1. The three main entities involved in grid experiments.

First of all, we need to realize that we must deal with the interaction of three main entities, as shown in Figure 1:

- **Component**: is in charge of carrying out one of the steps of the overall IR process. Examples of components which will be taken into consideration are:
  - *tokenizer*: processes documents and splits them into a stream of meaningful tokens;
  - *stop list*: removes stop words from the stream of tokens;
  - *stemmer*: reduces word forms to an approximation of a common morphological root;
  - *IR model*: defines the strategies according to which tokens are weighted and which matching functions have to be used;
  - *query construction*: defines the way in which queries have been created starting from topics;
  - *relevance feedback*: identifies the strategies for reformulating queries according to information provided (or assumed) about the results that are initially returned from a given query;
  - *cross-language & translation*: defines the techniques adopted for translating queries or documents from one language to another.
- **Language**: affects the performances and behaviour of the different components of an IRS due to its peculiar features, e.g. alphabet, syntax, morphology, and so on.
- **Task**: impacts the performances of IRS components due to its distinctive characteristics. Examples of tasks which will be taken into consideration are:
  - *monolingual*: topics and document collections are in the same language;
  - *bilingual*: the language of the topics (*source language*) is different from the language of the document collections (*target language*);
  - *multilingual*: topics in one source language have to be used for querying document collections in multiple target languages.

At first sight, we might think that only two entities would be enough to model the grid experiments, i.e. components and languages, since these are explicitly stated in the problem of understanding the behaviour of MLIA systems with respect to languages. Nevertheless, the notion of task is very important as well, because some of the components are very sensitive to task. So for example, query construction and relevance feedback change from monolingual to bilingual; the multilingual could theoretically also be involved, although the results are usually merged and this could smooth the impact of the tasks on the separate bilingual retrieval steps which are part of the overall multilingual retrieval. Furthermore, we should keep in mind that grid experiments are laboratory experiments and what works in grid experiments is not always the right decision for other kind of tasks, such as interactive retrieval or domain specific work.

Two issues have to be pointed about grid experiments before going into further details with their modelling, since they influence both they way in which the experiments have to be carried out and the way of interpreting and analysing the results.

Firstly, as depicted in Figure 1, the contributions of these three main entities to retrieval performances may be in general overlapping; nevertheless, at present, we do not have enough knowledge about this process to say whether, how, and how much these entities interact together, how much their contributions overlap – if they overlap – or how their contributions can be combined, e.g. in a linear fashion or according to some more complex relationships.

Secondly, the above issue partners with another long-standing problem in the IR experimentation: the impossibility of testing a single component without putting it within a complete IRS. In this respect, [5, p. 12] points out that "if we want to decide between alternative indexing strategies for example, we must use these strategies *as part of a complete information retrieval system*, and examine its overall performance (with each of the alternatives) directly". Therefore, we have to proceed by changing only one component at time and keeping all the others fixed, in order to point out the impact of that component on retrieval effectiveness; this also calls for the identification of suitable baselines with respect to which all the comparisons are made.

Definition 1: Let us define the following sets:

- C is a **set of components** and $c \in C$ is a generic **component**;
- L is a **set of languages** and $l \in L$ is a generic **language**;
- T is a **set of tasks** and $t \in T$ is a generic **task**;
- M is a **set of performance measures** and $m \in M$ is a generic **performance measure**.

We can introduce the **grid experiment** ge function as follows

$$\mathrm{ge} : C \times L \times T \rightarrow P \subseteq \mathbb{R}^{|M|}$$

where $p \in P$ are **performance measurements**, i.e. actual values, assigned to a triple (c, l, t) according to the performance measures defined in M and P is a **set of performance measurements**.

Definition 1 formalizes the notion of grid experiment we have discussed so far, models the fact that each performance measurement is obtained by the interaction of a component with a given language and task, and clarifies the name chosen for this type of experiments: in fact, we are going to create a kind of multi-dimensional table, i.e. a *grid*, where each cell is filled with performance measurements about a triple (c, l, t). Therefore, the grid experiments will experimentally determine the ge function; indeed, for each triple (c, l, t) in the domain, the grid experiments will determine the values, i.e. the performance measurements, to be associated to that triple in the codomain and computed according to some performance measure. Note that in most cases the codomain is $P = [0, 1]^{|M|}$ since usual performance measures, such as average precision, fall in this interval.

Furthermore, we should be aware that there might be implicit factors that influence our experimentation, besides the three main entities. For example, document collections impact the overall experimentation: different transliterations of the same language calls for some kind of normalization during tokenization and this may impact subsequent performances; the geographical area from which the document collections come may influence the retrieval process and the linguistic resources needed: German in Switzerland is different from German in Germany or Austria. Therefore, we should try to make these implicit factors as explicit as possible and keep them controlled or, at least, lower their impact on comparability of the results in order to avoid the introduction of a further entity in the design of the experiments.

Finally, we have to take into consideration the need for a proper presentation and visualization of the experimental results since this allows us both to better understand and study them and to more effectively communicate them to interested communities, as discussed in the next section.

III. Visualizing the Grid Experiment Function

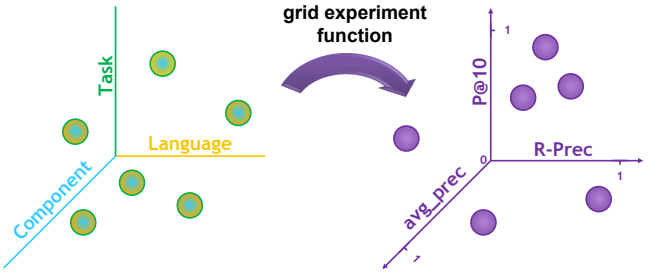

Fig. 2. Example of the grid experiment function.

Figure 2 shows an example of the grid experiment function which maps the three-dimensional domain into a subset the $[0, 1]$ cube, assuming that we have chosen only three performance measures to assess the experiments, e.g. average precision, R-precision, and P@10.

The example shows that if, on the one hand, the grid experiment function is a compact and easy way of modelling what grid experiments are, on the other hand its dimensionality may become too high and the visualization of the experimental results may become difficult and prevent a deep understanding of some interactions. Therefore, it should be possible to reduce the complexity of the grid experiment function by providing different *views* of it where its dimensionality, either in the domain or in the codomain, is reduced. For example, we could choose a given task and study the interaction between components and languages with respect to average precision; in this case, we could represent the experimental results in a three-dimensional cartesian space where for each pair (c, t) we can study the associated average precision value.

Moreover, the high dimensionality of the ge function poses some challenges also for the way of representing and visualizing it. In the following a couple of possible visualizations are discussed.

If we restrict ourselves to consider a single performance measure at time, we can represent the domain in a three-dimensional space and encode the actual values of the performance measure by using a color scale, as shown in Figure 3.

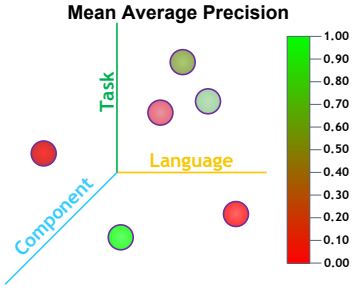

Fig. 3. Visualization of the grid experiment function one performance measure at time.

If we restrict ourselves to consider a single performance measure at time and only the interaction between components and languages, we can represent the domain in a bi-dimensional space and use the third dimension to plot the actual values of the performance measure, as shown in Figure 4.

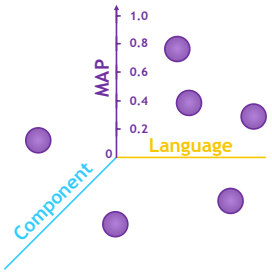

Fig. 4. Visualization of the grid experiment function one performance measure at time restricted to components and languages only.

The problem of a proper visualization of the experimental results is not a secondary issue since the outcomes of these experiments have to be shared not only with researchers of the IR field, who may be more used to deal with this material, but also with researchers and developers of relevant application communities, such as the DL one, who may be not very familiar with this stuff and may need some more

intuitive presentation of the results; some initial attempts in this direction has been carried out in [6], where a visual tool for comparing monolingual and bilingual performances has been proposed.

This issue also impacts the way in which grid experiments can promote the technology transfer, which will be discussed in more detail in the next section.

IV. Promoting Technology Transfer through Grid Experiments

A first way for promoting the knowledge transfer and favouring the creation of competencies about MLIA concerns the way in which grid experiments are organized.

Indeed, instead of conducting all the experiments within the TrebleCLEF consortium, some experiments could be assigned to external research groups of different nationalities in order to give them the possibility of dealing with multilinguality issues in a systematic way and distribute the work-load. This would also fit with one of the goals discussed in Section I: "[grid experiments could] help other countries to bring up their expertise in the IR field and create IR groups".

In order to ensure the robustness of this approach, we could proceed in two phases:

- *first phase*: is concerned with the consolidation of the methodology and the fixing of all the organizational aspects. It consists of selecting a core subset of components, languages, and task and running the experiments directly by the TrebleCLEF consortium. This will allow TrebleCLEF to gain experience on all the issue that may arise during the experiment and fine tune the overall procedure.
- *second phase*: is concerned with the involvement of the other research groups. It consists of organizing a "Grid Experiments Track" in the context the CLEF annual evaluation campaigns, where the rules for conducting experiments and properly describing them are very well defined and controlled, so that other research groups can participate in it. This would allow other research groups to participate in systematic investigation of the behaviour of an IRS with respect to languages by, perhaps, focussing on the subset of the grid experiments which concern their own language.

A second way of disseminating the knowledge gained with grid experiments is to use the *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* digital library system [7], [8] in order manage, make accessible, cite and reference, curate, enrich, and preserve the experiments. Moreover, DIRECT offer access not only to the experiments but also to performance measures and statistical analyses about them, as well as plots and reports summarizing them.

Moreover, a lot of care has been paid in the design and development of the DIRECT user interface in order to support high-level cognitive tasks and the investigation and understanding of the experimental results [9]. However, the visualization capabilities of DIRECT needs to be further extended in order to support the advanced three-dimensional visualizations discussed in the previous section.

V. Conclusions and Future Work

This paper provided an initial discussion of the issues related to the running of a series of grid experiments in the context of the TrebleCLEF project. In particular, we proposed an initial formalization which gave us the means for pointing out some relevant issues concerning them, we discussed how to effectively present and visualize the experimental results, and we presented how to promote the technology transfer about multilinguality to relevant target communities.

Future work will concern the careful selection of the actual components, languages, and task to take into consideration in the grid experiments as well as of the performance measures, descriptive statistics and hypothesis test needed to analyse them.

Acknowledgments

The reported work has been partially supported by the DELOS Network of Excellence on Digital Libraries, as part of the Information Society Technologies (IST) Program of the European Commission (Contract G038-507618). It has also been partially supported by the TrebleCLEF Coordination Action, as part of the Seventh Framework Programme of the European Commission, Theme ICT-1-4-1 Digital libraries and technology-enhanced learning (Contract 215231).

References

[1] M. Agosti, G. M. Di Nunzio, N. Ferro, D. Harman, and C. Peters, "The Future of Large-scale Evaluation Campaigns for Information Retrieval in Europe," in *Proc. 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2007)*, LNCS 4675, Springer, Heidelberg, Germany, 2007, pp. 509–512.

[2] M. Braschler, G. M. Di Nunzio, N. Ferro, J. Gonzalo, C. Peters, and M. Sanderson, "From CLEF to TrebleCLEF: promoting Technology Transfer for Multilingual Information Retrieval," in *Second DELOS Conference - Working Notes*, ISTI-CNR, Gruppo ALI, Pisa, Italy, December 2007.

[3] J. Savoy, "Report on CLEF-2001 Experiments: Effective Combined Query-Translation Approach," in *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross–Language Evaluation Forum (CLEF 2001) Revised Papers*, LNCS 2406, Springer, Heidelberg, Germany, 2002, pp. 27–43.

[4] C. W. Cleverdon, "The Cranfield Tests on Index Languages Devices," in *Readings in Information Retrieval*, K. Spärck Jones and P. Willett, Eds. Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA, 1997, pp. 47–60.

[5] S. E. Robertson, "The methodology of information retrieval experiment," in *Information Retrieval Experiment*, K. Spärck Jones, Ed. Butterworths, London, United Kingdom, 1981, pp. 9–31.

[6] F. Crivellari, G. M. Di Nunzio, and N. Ferro, "A Statistical and Graphical Methodology for Comparing Bilingual to Monolingual Cross-Language Information Retrieval," in *Information Access through Search Engines and Digital Libraries*, M. Agosti, Ed. Springer-Verlag, Heidelberg, Germany, 2008, pp. 171–188.

[7] M. Agosti, G. M. Di Nunzio, and N. Ferro, "The Importance of Scientific Data Curation for Evaluation Campaigns," in *Digital Libraries: Research and Development. First International DELOS Conference. Revised Selected Papers*, LNCS 4877, Springer, Heidelberg, Germany, 2007, pp. 157–166.

[8] G. M. Di Nunzio and N. Ferro, "DIRECT: a System for Evaluating Information Access Components of Digital Libraries," in *Proc. 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005)*, LNCS 3652, Springer, Heidelberg, Germany, 2005, pp. 483–484.

[9] M. Dussin and N. Ferro, "Design of the User Interface of a Scientific Digital Library System for Large-Scale Evaluation Campaigns," in *Second DELOS Conference - Working Notes*, ISTI-CNR, Gruppo ALI, Pisa, Italy, 2007.