Queries and Relevance Assessments: The Right Context for the Right Topic

Giorgio M. Di Nunzio[1] and Nicola Ferro[1]

Department of Information Engineering, University of Padua, Italy
{dinunzio, ferro}@dei.unipd.it

Abstract. We would like to discuss the problem of building a test collection for the evaluation of cross-language Information Retrieval (IR) systems. In particular, from the point of view of the experts that build the set of queries to test the performance system, and the assessors that judge the documents retrieved by the systems. Can the temporal and spatial context of a query and the user interaction history be a step forward to a more aware way to evaluate Cross-Language IR systems?

1 Introduction

The *Cross-Language Evaluation Forum (CLEF)* mainly aims at evaluating Cross-Language Information Retrieval systems that operate on multiple languages in both monolingual and cross-lingual contexts. The ad-hoc track in CLEF adopts a corpus-based, automatic scoring method for the assessment of system performance, based on ideas first introduced in the Cranfield experiments in the late 1960s. The test collection used consists of a set of "topics" describing information needs and a collection of documents to be searched to find those documents that satisfy these information needs. Evaluation of system performance is then done by judging the documents retrieved in response to a topic with respect to their relevance, and computing the recall and precision measures. The distinguishing feature of CLEF is that it applies this evaluation paradigm in a multilingual setting. This means that the criteria normally adopted to create a test collection, consisting of suitable documents, sample queries and relevance assessments, have been adapted to satisfy the particular requirements of the multilingual context.

2 The Right Context for the Right Topic

Given the experience gained being the research group responsible for the management of the CLEF technical infrastructure, we would like to bring to your attention two problems: building the set of topics, and the set of relevance judgements, in a multilingual context. In particular, would it be sensible to apply Adaptive Information Retrieval techniques for the creation of the set of queries and relevance assessments?

The creation of a set of queries suitable for a certain kind of task (ad-hoc retrieval, domain specific retrieval, geographical retrieval) is a long process. This

process requires the effort of a group of experts that have to find the right set of queries that are neither too general nor too specific; moreover, in a multilingual environment, each query should find answers also in collections of documents written in different languages and that cover different time intervals. In order to overcome this problem, the set of queries used this year in the ad-hoc track of CLEF were split into two subsets: a set of *general* queries, i.e. answers can be found in different years and different geographical locations, and a set of *specific* queries, i.e. queries that are strictly coupled with a specific collection of document and language. This fact suggests that each query has an implicit, or explicit as in this case, geographical temporal context; this context can be used to help the experts to understand whether a particular formulation of a topic is suitable or not. The idea of the context and adaptation to user behavior and experience is even more founded when you think at the process of building a query as an interactive process that requires user's feedback to an IR system in order to tune the difficulty of the query.

A similar consideration could be done for the relevance assessments. The act of judging the relevance of a subset of the documents retrieved by a system given a topic requires the assessors to scan a long list of documents. In this task human abilities and experience play an important role. The assessors of the CLEF wanted the buttons of the relevance assessment interface placed in such a way to assess as fast as possible. However, the process is so long that there is a strict limit on the number of documents that can be judged for each language. If you consider that only a few hundreds of documents are relevant over some tens of thousands, it would be vital for the assessors to rapidly focus their effort only on relevant documents. In this sense, a user interaction history, that creates the context for each particular query, may be used to skip non–relevant documents and read relevant ones only.

Solutions to these problems may be found in the use of systems like MIRACLE[1] designed for interactive Cross-Language Information Retrieval, or the use of implicit relevance feedback models like those ones presented in[2], or techniques like the Interactive Searching and Judging (ISJ) method tested by[3], or new approaches of considering the evaluation campaigns data as scientific data to be cured in order to support in-depth evaluation[4].

References

1. He, D., Oard, D.W., Wang, J., Luo, J., DemnerFushman, D., Darwish, K., Resnik, P., Khudanpur, S., Nossal, M., Subotin, M., and Leuski, A. Making MIRACLEs: Interactive Translingual Search for Cebuano and Hindi, ACM Transactions on Asian Language Information Processing (TALIP) **2**(3) (2003) 219–244
2. White, R.W., Ruthven, I., Jose, J.M., and Van Rijsbergen, C. J., Evaluating Implicit Feedback Models Using Searcher Simulations, ACM Transactions on Information Systems **23**(3) (2005) 325–361
3. Sanderson, M. and Joho, H.: Forming Test Collections with No System Pooling. Proceedings of the SIGIR 2004, (2004) 33-40 .
4. Agosti, M., Di Nunzio G.M., Ferro, N.: A Data Curation Approach to Support In-depth Multilingual Evaluation Studies, Proceedings of the Workhshop on New Directions in Multilingual Information Access (MLIA at SIGIR'06), (2006) 65–68