Large-Scale Interactive Evaluation of Multilingual Information Access Systems – the iCLEF Flickr Challenge

[1]Paul Clough, [2]Julio Gonzalo, [3]Jussi Karlgren, [1]Emma Barker, [2]Javier Artiles, [2]Victor Peinado
[1]University of Sheffield, UK
[2]Universidad Nacional de Educación a Distancia, Spain
[3]Swedish Institute of Computer Science, Sweden

ABSTRACT

Participation in evaluation campaigns for interactive information retrieval systems has received variable success over the years. In this paper we discuss the large-scale interactive evaluation of multilingual information access systems, as part of the Cross-Language Evaluation Forum evaluation campaign. In particular, we describe the evaluation planned for 2008 which is based on interaction with content from Flickr, the popular online photo-sharing service. The proposed evaluation seeks to reduce entry costs, stimulate user evaluation and encourage greater participation in the interactive track of CLEF.

1. EVALUATION OF IR SYSTEMS

Evaluating the performance of Information Retrieval (IR) systems is an important part of the system development process from an engineering point of view, and a crucial part of the research process. It enables development of useful and effective technology, together with generalisable and sustainable knowledge for future development cycles. A systematic, transparent, and intuitively valid evaluation process has been a defining and unifying feature of the information access research field during the past decades ([14][15][16][6][3]), and has been instrumental in ensuring simultaneous commercial success and academic stringency. We should stay true to this tradition.

1.1. Traditional Evaluation Methodologies

The evaluation of retrieval systems tends to focus on either the system or the user. Saracevic [14] distinguishes six levels of evaluation for information systems that include information retrieval systems: (1) at the engineering level, (2) at the input level, (3) at the processing level, (4) at the output level, (5) at the use and user level and (6) at the social level. For many years information access evaluation has tended to focus on the first three levels, predominately through the use of standardized benchmarks (or reference collections) in a laboratory-style setting (also known as batch-mode evaluation). The Cranfield experiments [5] were some of the first to develop and demonstrate the use of lab-based evaluation. However, information access systems are most commonly used interactively, within a task and social context, and this drives the need for user-centered evaluation to address performance at the latter three levels (output, use and user, and social). User-centered evaluation is important because it assesses the overall success of a retrieval system (as determined by end users of the systems) which takes into account other factors other than system performance, e.g. task context, cognitive influence, and the design of the user interface (see, e.g. [8]).

To enable reproducibility and comparison, standardized resources for evaluating document retrieval systems have been designed and used (a.k.a. test collections) for at least 30 years (first proposed in the Cranfield I and II projects [4]). Standardized resources have been used in major information access evaluation campaigns around the world such as TREC[1], CLEF[2] and NTCIR[3]. Researchers have recognized the value of testing retrieval systems within the large-scale setting through organized and managed campaigns, undoubtedly acting as a major influence in the design of information access systems over the past ten years or so. Not only have these events provided a testbed for evaluation, but also an interactive forum in which to exchange ideas and discuss techniques for successful system and algorithm design.

Although primarily a testbed for system-orientated evaluation, these campaigns (in particular TREC and CLEF) have also included user-oriented (or interactive) evaluation. However, evaluating interactive information access systems experimentally is challenging [2][7]. The high effort, cost, and overhead involved in recruiting test subjects, designing test systems, and formulating experimental scenarios risks both delivering unrealistic laboratory-based task formulations, and finding general results drowned in inter-user variation. The low reproducibility of experiments, failure to effectively generalize results, and the difficulty of comparison between different systems has limited the success of such initiatives (see, e.g. [7][12]).

1.2. The Challenge for Interactive Evaluation

However successful evaluation schemes have been in the past, new media pose challenges to content analysis and to established target notions of "relevance"; new modes of communication and contexts pose challenges to use cases and tasks underlying traditional ad-hoc evaluation schemes; multilingual materials, audience, and usage situations pose challenges to systems and processing resources. In addition, new interactive services are taken up by user communities, not by virtue of their engineering qualities or their ergonomics but by consumer evaluation based on social factors, marketing effectiveness, or even legal requirements: offering a well-built interface and providing solid content is no guarantee to commercial success. Evaluating interactive retrieval must make itself relevant to service providers by evaluating those aspects of

[1] http://trec.nist.gov/ [accessed 11/03/08]
[2] http://www.clef-campaign.org/ [accessed 11/03/08]
[3] http://research.nii.ac.jp/ntcir/ [accessed 11/03/08]

interaction that are most crucial for the task a system is designed for: if the system has no underlying task model it must acquire one to be valuable. Traditional ad-hoc evaluation schemes have had an implicit use case and task model which does not necessarily carry over to new situations.

The next generation of evaluation methodologies must take into account not only changes in the underlying content, but the varying user base and societal and contextual factors surrounding the usage under study. How might we find a task that allows us to evaluate interactive retrieval, using multi-medial and multilingual data, possibly not in a standard collection, affording the potential to model new settings, new contexts, new tasks with large enough numbers of users to transcend inter-user noise, with a minimal amount of administrative overhead, and yet provide generalisable, intellectually appealing, and potentially interesting and useful results?

2. EVALUATING MULTILINGUAL IR

Multilingual information retrieval (MLIR) describes the situation in which a user searches for information in a language different from the query (see, e.g. [9]). Multilingual information retreival can be thought of as a combination of machine translation and traditional monolingual information retrieval. Most research has focused on locating and exploiting translation resources with which the user's search requests or target documents (or both) are translated into the same language. Campaigns such as the Cross Language Evaluation Forum (CLEF) [13)] and the Text REtrieval Conference (TREC) [2][17] multilingual track have helped encourage and promote international research, as well as create standardised resources for evaluating multi-lingual information access approaches.

2.1. Interactive CLEF (iCLEF)

The CLEF interactive track (iCLEF[4]) has been devoted, since 2001, to the study of Cross-Language Information Retrieval from a user-centered perspective. The aim has always been to investigate real-life cross-language searching problems in a realistic scenario, and to obtain indications on how best to aid users in solving them (see, e.g. [11]). Multilingual information retrieval is particularly interesting from an interactive point of view, because the need for search assistance is substantially higher than in monolingual information retrieval: normally, the user can quickly adapt to the system's modus operandi, but not to an unknown target language.

iCLEF experiments have investigated the problems of foreign-language text retrieval, question answering and image retrieval, including aspects such as query formulation, translation and refinement, document selection and document examination. The focus has always been on improving the outcome of the process in terms of a classic notion of relevance (documents meeting an information need that prompted a query), and the target collection (except for image search experiments) has always consisted of news texts in languages foreign to the user. Finally, the task has always involved the comparison of a reference system with a contrastive system, combining users, topics and systems with a Latin-Square design to detect system effects and filter out other effects (as used within the Interactive TREC track [7]).

[4] http://nlp.uned.es/iCLEF/ [accessed 11/03/08]

Table 1: iCLEF task goals and participation (2001-2006).

Year	Task	Goal	Groups
2001	Ad-hoc	Document selection	3
2002	Ad-hoc	Document selection, query formulation & reformulation	5
2003	Ad-hoc	Full Cross-Language search	5
2004	QA	Full Cross-Language QA	5
2005	Image search/QA	Full Cross-Language QA / known-item image search	5 (2 image; 3 QA)
2006	Image search	open	3

Table 1 shows the progression of iCLEF since 2001. Overall, participation has always been low, with a high of 5 participating groups; a low of 3 groups. Although iCLEF in only a few years of activity has established the largest collected body of knowledge on the topic of interactive cross-language information retrieval, the experimental setup has proven limited in certain respects:

- The search task itself is unrealistic: news collections are comparable across languages, and most of the pertinent information tends to be available in the user's native language. Therefore, why would a user search for this information in an unknown language?
- The target notion of "relevance" does not cover all aspects that make an interactive search session successful (e.g. other factors could include satisfaction of results, usability of the interface itself, and the system's response time).
- The Latin-Square design imposes heavy constraints on the experiments, making them costly and with a limited validity (the number of users is necessarily limited, and statistically significant differences are hard to obtain).

2.2. Moving to Flickr

In order to overcome these limitations, the iCLEF track moved to a new pilot framework in 2006 [4] [10]: we decided to use the publicly available (and immensely popular) photo-sharing service Flickr[5] as the target collection. This is an inherently multi-lingual database through its lively tagging and commenting features, and it has the potential to offer a range of challenging and realistic multilingual search tasks for interactive experimentation. Although the database is in constant evolution – something which compromises reproducibility – the Flickr search API allows specifying timeframes (e.g. search images uploaded in the period 2004-2007), which permits defining a more stable dataset for experiments.

2.3. The Experience of iCLEF2006

Besides moving to Flickr as the target database, in 2006 we took the following additional decisions:

1. To lower the threshold of entry to the evaluation campaign, we offered a standard multi-lingual interface which various

[5] http://www.flickr.com [accessed 11/03/08]

research sites can use to explore whatever features of interaction they are most interested in. The interface provides a (baseline) term translation service and a fine-grained log of user actions.

2. We designed three different search tasks: known-item search (find this image), topical search (find as many pictures as possible around this topic), and text illustration (find good images to illustrate this text). The illustration task naturally provides a search scenario where evaluation has to go beyond the traditional notion of topical relevance.
3. We did not impose any evaluation methodology on the participants. Being a novel evaluation scenario, we wanted to involve iCLEF participants in the exploration of novel evaluation methodologies as a key part of the campaign. This made the 2006 a collaborative exercise on how to study interactive issues in cross-lingual multi-medial information access.

Whilst we found enthusiastic support from the potential participants (fourteen groups signed up for the task), only three sites actually participated in the final evaluation (the three organizing groups themselves). We found that while the freedom of the task appeared to be attractive at first sight, the entry threshold was still too high: building an interface and designing an experiment proved too costly and the open design provided too little support for newcomers. In addition, we found that the submission schedule used in other CLEF tracks collapses with iCLEF due to the inherent time-consuming nature of implementing a user interface and running interactive experiments.

As in previous iCLEF editions, there was valuable knowledge acquired, but little participation from the research community. It can be concluded that, similar to Interactive TREC, the interactive CLEF task has not been as successful as the lab-based system-orientated tasks. Possible reasons for this include:

- Considering users is just not seen as important in information retrieval evaluation (compared to system-oriented evaluation).
- The large-scale setting of an evaluation campaign is simply ill-suited to interactive evaluation.
- Performing user experiments is time-consuming and difficult and little gain is seen for it (e.g. lack of generality and difficulty in comparing results).
- Developing efficient algorithms for information access is considered more important than user-orientated issues.
- System-orientated is well-understood; user-orientated evaluation is less clear and requires a deeper understanding (e.g. in the experimental design).

2.4. Remedies for iCLEF2008

One of the main limitations of iCLEF 2006 was that, although we moved into a realistic multilingual search setting, the experiment designed still did not facilitate having large-scale user logs. All three experiments employed less than 30 users that had to be recruited, trained, monitored and controlled. In 2008 we decided to concentrate on collecting user logs at a larger scale, and let participants concentrate on mining such logs to gain more knowledge about how users behave when they need to search in unfamiliar languages.

To be able to harvest a substantially larger set of search sessions, we decided to implement a single, basic multilingual search interface for Flickr, and make it available in the web for anyone. To attract – and specially to keep - potential users, we have made the search task a game. The basic task is simple: finding a given image (the user is shown a picture) in Flickr. Finding more images improves the user ranking in a "Hall of Fame". Note that this is a fully multilingual task: the image to be found can be annotated in any (or several) of the target languages, and the user does not know a priori which is the case.

This was modeled on the success of the ESP game for labeling images [1] and thought to increase interest in the task for both participating groups and their subjects. The entry costs of iCLEF2006 were clearly still too high, therefore for 2008 we provide groups with an experimental design, but still allow open extension for groups to adapt the design for their own investigations. As the evaluation has moved to Flickr/Web users, participants now have something in common with the subjects they recruit, therefore are more likely to be a captive set of subjects. Finally, to allow for the timing differences of running an interactive evaluation task, we have adjusted the deadlines of the standard iCLEF calendar, giving participants more time to run and analyse their experiments.

3. THE iCLEF2008 TRACK

We now describe the iCLEF track for 2008 in terms of what the organizers provide to participating groups, and what the groups must do.

3.1. Data and Resources

The organizers of iCLEF are providing the following to participants in 2008:

3.1.1. Task definition

The task for 2008 is known-item image retrieval based on photos from Flickr: the user is given an image, and the goal for them is to find the image again from Flickr. The advantage of this kind of search task is that it has clear goals for the user, it has a clearly defined measure of success (the image is either found or not) and whilst searching for the required image, users will invoke different (and potentially interesting) search patterns. The user does not know in advance in which languages the image is annotated; therefore searching in multiple languages is essential to successfully find the images. The task is organised as a game: the more images found, the higher users (and user groups) will be ranked. Section 3.3 describes in more detail the selection of topics and example images.

3.1.2. Default MLIR front-end to Flickr

We have designed and implemented a multilingual information retrieval interface to Flickr with the following functionalities (shown in Figure 1.):

- Multilingual search: query in one language, get search results in up to six languages (English, Spanish, French, Italian, Dutch and German).
- Term-to-term translations between six languages (English, Spanish, German, French, Dutch and Italian) using freely available dictionaries (taken from http://xdxf.revdanica.com/down/).

- Selection of "best" target translations according to (i) presence in the Flickr *related terms* for the query, which often include target-language terms because they co-occur with the query terms in images annotated in multiple languages, something which is not unusual in the Flickr database; and (ii) string similarity between the source and target words. This was included because the free dictionaries used did not have information about the most frequent sense/translation.
- Enables user to pick/remove translations, and add their own translations (which go into a "personal dictionary"). We did not provide back-translations to support this process, in order to study correlations between target language abilities (active, passive, none) and selection of translations.
- Provision of search suggestions (Flickr related terms plus tags from displayed images).
- Control over the game-like features of the task: flow of images, users ranking, etc.

Note that we did not intend to provide the best possible cross-language assistance to search the Flickr collection. Our intention was to provide a rather standard, baseline interface where we can get information from users' behavior which is not too much dependent on a particular interface idiosyncrasy.

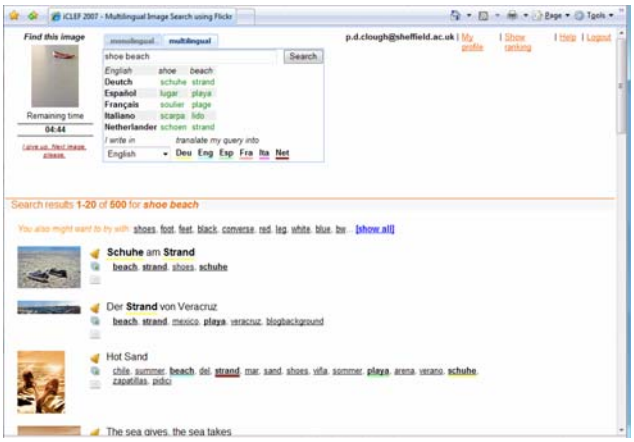

Figure 1: The iCLEF2008 interface.

3.1.3.Experiment customization

In addition to harvesting search logs, we also offer this interface for groups interested in performing their own experiments with selected types of users, and we provide support for customization of the interface.

3.1.4.Generation of search logs

Search logs will be generated from the interface. We will focus on two user groups: (i) CLEF participants, which will be asked to play the Flickr game (the best team will receive an award at CLEF 2008), and (ii) Flickr/Web users at large. The game will be publicized in order to get a substantial amount of usage information.

The idea of using CLEF researchers as a user group is not simply a matter of convenience: we believe that few cross-lingual information retrieval researchers have actually experienced cross-language search tasks as users, and the exercise we propose might broaden their vision of cross-lingual information retrieval research.

3.2. Participating in the Track

Participants in iCLEF2008 can essentially do two tasks: analyse log files based on all participating users (which is the default option) and perform their own interactive experiments with the interface provided by the organization. CLEF individuals will register in the interface as part of a team, so that a ranking of teams can be produced in addition to a ranking of individual users.

3.2.1 Generation of search logs

Participants can mine data from the search session logs, for example looking for differences in search behaviour according to language skills, or correlations between search success and search strategies.

3.2.2 Interactive experiments

Participants can recruit their own users and conduct their own experiments with the interface. For instance, they could recruit a set of users with passive language abilities and another with active abilities in certain languages and, besides studying the search logs, they could perform observational studies on how they search, conduct interviews, etc.

3.3. Topic Selection

In total, 180 example images will be available within the system for users to find (30 images in each language set: German, Spanish, English, French, Italian and Dutch). Classification of the language of an image is based on the "main" language of an image's text and tagset. Rather than select images randomly from Flickr, we wanted to maintain some element of experimental control and topic variation. The following points were considered during selection of the images:

- There should be sufficient text/tags accompanying an image to facilitate the task (i.e. we required "rich" text where possible).
- Ideally we wanted diverse topics in the test set and required roughly equivalent subject/topics in the different language groups, so the aim was to get at least one instance of a subject/topic group, for each of the language sets.
- When collecting images in different languages but with the same subject/topic, we aimed to find images with a similar visual perspective.
- The known item task must not be too hard: queries for finding images were manually recorded and an independent search carried out to check the images are not too hard to find.

Figure 2 shows example images from the current set of topics. As can be seen, these vary in aspects such as subject (topical content of an image), visual content, orientation, activity depicted in the image, and visual perspective (e.g. close-up, long distance).

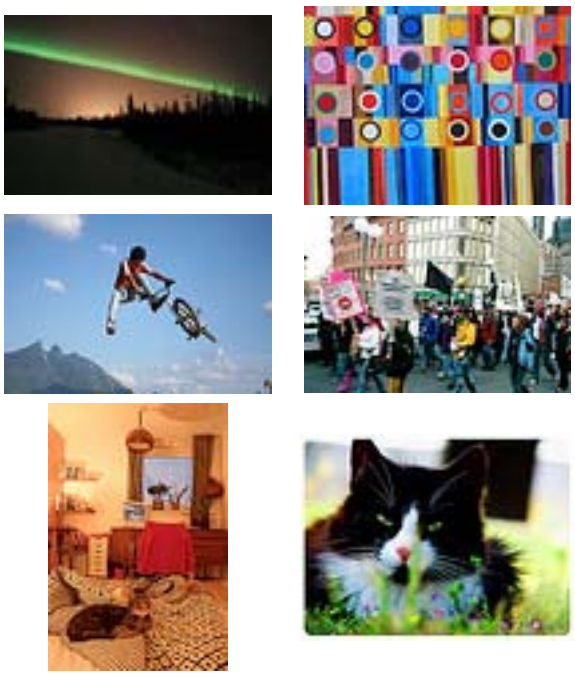

Figure 2: Example topics for known-item search.

4. CONCLUSIONS

The iCLEF task has so far provided a substantial body of knowledge around the interactive aspects of Cross-Language Retrieval, but it has failed to engage the cross-lingual information retrieval research community, and it has always been restricted to experiments with a limited set of users, where statistically significant insights are hard to find. In the design of iCLEF 2008 we have made a significant change in our experiment design, focusing on acquiring a large set of search session logs and offering the data to iCLEF participants, so that the task focus is on mining search logs rather than designing interactive experiments. At the same time, we have decided to engage the CLEF research community as a user group for the experiment, hoping that this fully multilingual search exercise will broaden the scope of midstream cross-lingual information retrieval research into the essential – but hard to study systematically – interactive aspects of multilingual retrieval.

ACKNOWLEDGEMENTS

Work partially funded by the TrebleCLEF Coordination Action (FP7-ICT-2007-1).

5. REFERENCES

[1] von Ahn, L. Games with a Purpose, Computer, Vol. 39(6), pp. 92-94, June, 2006.

[2] Belkin, N.J., Dumais, S.T., Scholtz, J. & Wilkinson R. Evaluating interactive information retrieval systems: opportunities and challenges. CHI Extended Abstracts 2004: 1594-1595.

[3] Buckley, C. & E. M. Voorhees. Retrieval system Evaluation. In E. M. VOORHEES & HARMAN D. K. (Eds.), TREC: experiment and evaluation in information retrieval. London, England, MIT Press. 2005.

[4] Clough, P., Gonzalo, J. & Karlgren, J. Multilingual interactive experiments with Flickr. EACL 2006 Workshop on New Text - Wikis and blogs and other dynamic text sources. 2006.

[5] Harter, S. P. & Hert, C. A. Evaluation of information retrieval systems: Approaches, issues, and methods. Annual Review of Information Science and Technology (ARIST), 32, 3-94. 1997

[6] Hersh, W. Information Retrieval: A Health and Biomedical Perspective, 2nd edition, Springer-Verlag: New York, Berlin, Heidelberg, 2005.

[7] Hersh, W. & Over, P. Interactivity at the Text Retrieval Conference (TREC) Information Processing & Management, Volume 37, Issue 3, May 2001, Pages 365-367.

[8] Ingwersen, P. & Järvelin, K. The turn: integration of information seeking and retrieval in context, Springer. 2005.

[9] Jones, G.J.F., Beyond English Text: Multilingual and Multimedia Information Retrieval, in Charting a New Course: Natural Language Processing and Information Retrieval. Essays in Honour of Karen Sparck Jones (ed. J. Tait), 2005, Kluwer. 2005.

[10] Karlgren, J., Gonzalo, J. & Clough, P. iCLEF 2006 Overview: Searching the Flickr WWW Photo-Sharing Repository. In CLEF 2006 Proceedings. 2007.

[11] Oard, D. & Gonzalo, J. The CLEF 2003 Interactive Track, Comparative Evaluation of Multilingual Information Access Systems. Results of the CLEF 2003 Evaluation Campaign. Springer-Verlag LNCS 3237, 2004.

[12] Over, P. The TREC interactive track: an annotated bibliography Information Processing & Management, Volume 37, Issue 3, May 2001, Pages 369-381.

[13] Peters, C. and Braschler, M.: Cross Language System Evaluation: The CLEF Campaigns. Journal of the American Soc. for Inf. Sci. and Tech. Vol. 52(12) (2001) 1067-1072.

[14] Saracevic, T. 1995. Evaluation of evaluation in information retrieval. In Proceedings of the 18th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Seattle, Washington, United States, July 09 - 13, 1995). E. A. Fox, P. Ingwersen, and R. Fidel, Eds. SIGIR '95. ACM Press, New York, NY, 138-146.

[15] Spark Jones, K. (Ed.). 1981. Information Retrieval Experiment. London: Butterworths.

[16] Spark Jones, K. & Willett, P. (Eds.). 1997. Readings in Information Retrieval, San Francisco, CA: Morgan. Kaufmann Publishers, Inc.

[17] Voorhees, E.M. & Harman, D.: Overview of TREC 2001, In NIST Special Publication 500-250: Proceedings of TREC2001, NIST, 2001.