Evaluating Interactive Cross-Language Information Retrieval: Document selection

Douglas W. Oard

Human Computer Interaction Laboratory
College of Information Studies and
Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742, USA
oard@glue.umd.edu.edu,
WWW home page: http://www.glue.umd.edu/~oard/

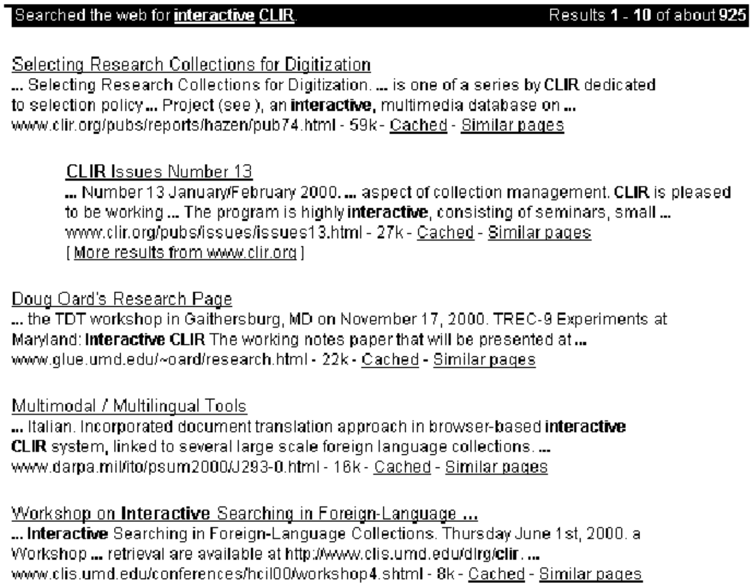
Abstract. The problem of finding documents that are written in a language that the searcher cannot read is perhaps the most challenging application of Cross-Language Information Retrieval (CLIR) technology. The first Cross-Language Evaluation Forum (CLEF) provided an excellent venue for assessing the performance of automated CLIR techniques, but little is known about how searchers and systems might interact to achieve better cross-language search results than automated systems alone can provide. This paper explores the question of how interactive approaches to CLIR might be evaluated, suggesting an initial focus on evaluation of interactive document selection. Important evaluation issues are identified, the structure of an interactive CLEF evaluation is proposed, and the key research communities that could be brought together by such an evaluation are introduced.

1 Introduction

Cross-language information retrieval (CLIR) has somewhat uncharitably been referred to as "the problem of finding people documents that they cannot read." Of course, this is not strictly true. For example, multilingual searchers might want to issue a single query to a multilingual collection, or searchers with a limited active vocabulary (but good reading comprehension) in a second language might prefer to issue queries in their most fluent language. In this paper, however, we focus on the most challenging case—when the searcher cannot read the document language at all.

Before focusing on evaluation, it might be useful to say a few words about why anyone might want to find a document that they cannot read. The most straightforward answer, and the one that we will focus on here, is that after finding the document they could somehow obtain a translation that is adequate to support their intended use of the document (e.g., learning from it, summarizing it, or quoting from it). CLIR and translation clearly have a symbiotic relationship—translation makes CLIR more useful, and CLIR makes translation more useful (if you never find a document that you cannot read, why would you need translation?).

In the research literature, it has become common to implicitly treat CLIR as a task to be accomplished by a machine. Information retrieval is a challenging problem, however, and many applications require better performance than machines alone can provide. In such cases, the only practical approach is to develop systems in which humans and machines interact to achieve better results than a machine can produce alone. A simple example from monolingual retrieval serves to illustrate this point. Figure 1 shows the result of a Google search for "interactive CLIR." The top-ranked documents are about interactive products developed by the Council on Library and Information Resources. But an interactive searcher can easily recognize from the brief summaries that the next few documents in the ranked list are on topic. In this case, a system that might be judged a failure if used in a fully automatic (top-document) mode actually turns out to be quite useful when used as the automatic portion of a human-machine system.



Searched the web for **interactive** CLIR. Results **1** - **10** of about **925**

Selecting Research Collections for Digitization
... Selecting Research Collections for Digitization. ... is one of a series by **CLIR** dedicated to selection policy ... Project (see), an **interactive,** multimedia database on ...
www.clir.org/pubs/reports/hazen/pub74.html - 59k - Cached - Similar pages

CLIR Issues Number 13
... Number 13 January/February 2000. ... aspect of collection management. **CLIR** is pleased to be working ... The program is highly **interactive**, consisting of seminars, small ...
www.clir.org/pubs/issues/issues13.html - 27k - Cached - Similar pages
[More results from www.clir.org]

Doug Oard's Research Page
... the TDT workshop in Gaithersburg, MD on November 17, 2000. TREC-9 Experiments at Maryland: **Interactive CLIR** The working notes paper that will be presented at ...
www.glue.umd.edu/~oard/research.html - 22k - Cached - Similar pages

Multimodal / Multilingual Tools
... Italian. Incorporated document translation approach in browser-based **interactive CLIR** system, linked to several large scale foreign language collections. ...
www.darpa.mil/ito/psum2000/J293-0.html - 16k - Cached - Similar pages

Workshop on **Interactive** Searching in Foreign-Language ...
... **Interactive** Searching in Foreign-Language Collections. Thursday June 1st, 2000. a Workshop ... retrieval are available at http://www.clis.umd.edu/dlrg/**clir**. ...
www.clis.umd.edu/conferences/hcil00/workshop4.shtml - 8k - Cached - Similar pages

Fig. 1. Top Google search results for "interactive CLIR."

The process by which searchers interact with information systems to find documents has been extensively studied (for an excellent overview, see [3]). Essentially, there are two key points at which the searcher and the system interact: query formulation and document selection. Query formulation is a complex cognitive process in which searchers apply three kinds of knowledge—what they think they want, what they think the information system can do, and what they think the document collection being searched contains—to develop a query. The query formulation process is typically iterative, with searchers learning about the collection and the system, and often about what it is that they really wanted to know, by posing queries and examining retrieval results. Ultimately we must study the query formulation process in a cross-language retrieval environment if we are to design systems that effectively support real information seeking behaviors. But the Cross-Language Evaluation Forum (CLEF) is probably not the right venue for such a study, in part because the open-ended nature of the query formulation process might make it difficult to agree on a sharp focus for quantitative evaluation in the near term.

Evaluation of cross-language document selection seems like a more straightforward initial step. Interactive document selection is essentially a manual detection problem—given the documents that are nominated by the system as being of possible interest, the searcher must recognize which documents are truly of interest. Modern information retrieval systems typically present a ranked list that contains summary information for each document (e.g., title, date, source and a brief extract) and typically also provide on-demand access to the full text of one document at a time. In the cross-language case, we assume that both the summary information and the full text are presented to the searcher in the form of automatically generated translations—a process typically referred to as "machine translation."[1] Evaluation of document selection seems to be well suited to the CLEF framework because the "ground truth" needed for the evaluation (identifying which documents *should have* been selected) can be determined using the same pooled relevance assessment methodology that is used in the present evaluation of fully automatic systems

Focusing on interactive CLIR would not actually be as a radical departure for CLEF as it might first appear. As Section 2 explains, the principal CLEF evaluation measure—mean average precision—is actually designed to model the automatic component of an interactive search process, at least when used in a monolingual context. Section 3 extends that analysis to include the effect of document selection, concluding that a focused investigation of the cross-language document selection problem is warranted. Sections 4 and 5 then sketch out the broad contours of what an interactive CLEF evaluation with such a focus might look like. Finally, Section 6 addresses the question of whether the necessary research base exists to justify evaluation of interactive CLIR by identifying some key research communities that are well positioned to contribute to the development of this technology.

[1] Note that the subsequent translation step—translation to support the ultimate use of the document—may or may not be accomplished using machine translation, depending on the degree of fluency that is required.

2 Deconstructing Mean Average Precision

Two types of measures are commonly used in evaluations of cross-language information retrieval effectiveness: ranked retrieval measures and set-based retrieval measures. In the translingual topic tracking task of the Topic Detection and Tracking evaluation, a set based measure (detection error cost) is used. But ranked retrieval measures are reported far more commonly, having been adopted for the cross-language retrieval tasks in CLEF, TREC and NTCIR. The `trec_eval` software used in all three evaluations produces several useful ranked retrieval measures, but comparisons between systems are most often based on the mean uninterpolated average precision (MAP) measure. MAP is defined as:

$$MAP = E_i[E_j[\frac{j}{r(i,j)}]$$

where $E_i[\ ]$ is the sample expectation over a set of queries, $E_j[\ ]$ is the sample expectation over the documents that are relevant to query i, and $r(i,j)$ is the rank of the j^{th} relevant document for query i.

The MAP measure has a number of desirable characteristics. For example, improvement in precision at any value of recall or in recall at any value of precision will result in a corresponding improvement in MAP. Since MAP is so widely reported, it is worth taking a moment to consider what process the computation actually models. One way to think of MAP is as a measure of effectiveness for the one-pass interactive retrieval process shown in Figure 2 in which:

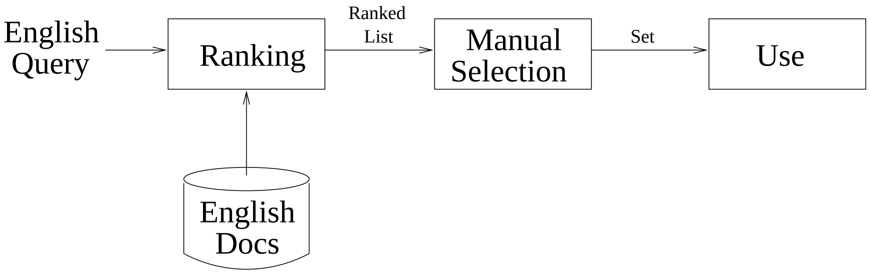


Fig. 2. A one-pass monolingual search process.

1. The searcher creates a query in a manner similar to those over which the outer expectation is computed.

2. The system computes a ranked list in a way that seeks to place the topically relevant documents as close to the top of the list as is possible, given the available evidence (query terms, document terms, embedded knowledge of language characteristics such as stemming, . . .).

3. The searcher starts at the top of the list and examines each document (and/or summaries of those documents) until they are satisfied.

4. The searcher becomes satisfied after finding some number of relevant documents, but we have no *a priori* knowledge of how many relevant documents it will take to satisfy the searcher. Note that here we implicitly assume that every document is either relevant or it is not (in other words, we don't account for differences in the perceived degree of relevance), and that relevance assessments are independent (i.e., having seen one document does not change the searcher's opinion of the relevance of another relevant document).

5. The searcher's degree of satisfaction is related to the number of documents that they need to examine before finding the desired number of relevant documents.

Although actual interactive search sessions often include activities such as learning and iterative query reformulation that are not modeled by this simple process, it seems reasonable to expect that searchers would prefer systems which perform better by this measure over systems that don't perform as well.

3 Modeling the Cross-Language Retrieval Process

One striking feature of the process described above is that we have implicitly assumed that the searcher is able to recognize relevant documents when they see them. Although there will undoubtedly be cases when a searcher either overlooks a relevant document or initially believes a document to be relevant but later decides otherwise, modeling the searcher as a perfect detector is not an unreasonable assumption when the documents are written in a language that the searcher can read. If the documents are written in a language that the searcher can not read, the final three steps above could be modified as illustrated in Figure 3 to:

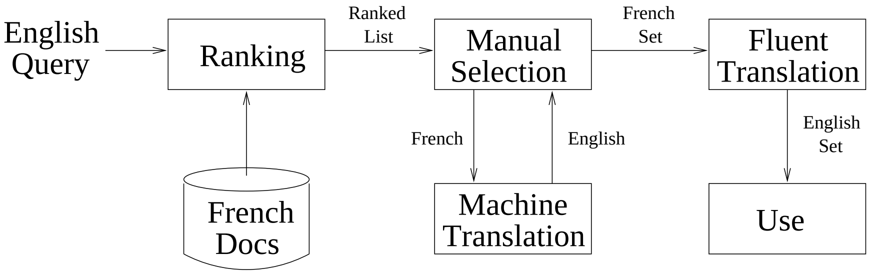


Fig. 3. A one-pass cross-language search process for searchers who cannot read French.

3a. The searcher starts at the top of the list and examines **an automatically produced translation** of each document (**or summary translations** of those documents) until they are satisfied.

4.a The searcher becomes satisfied after identifying a number of **possibly relevant** documents that **they believe is sufficient** to assure that they have found the desired number of relevant documents, but we have no *a priori* knowledge of how many relevant documents it will take to satisfy the searcher. [2]

5a. The searcher commissions **fluent translations of the selected documents**, and the searcher's degree of satisfaction is related to both the number of documents that they needed to examine and the fraction of the translated documents that actually turn out to be relevant.[3]

Of course, this is only one of many ways in which a cross-language retrieval system might be used.[4] But it does seem to represent at least one way in which a cross-language retrieval system might actually be employed, and it does so in a way that retains a clear relationship to the MAP measure that is already in widespread use. The actual outcome of the process depends on two factors:

– The degree to which the automatically produced translations support the searcher's task of recognizing possibly relevant documents.
– The searcher's propensity to select documents as being possibly relevant in the presence of uncertainty.

We model the combined effect of these factors using two parameters:

p_r The probability of correctly recognizing a relevant document.
p_f The probability of a false alarm (i.e., commissioning a translation for a document that turns out not to be relevant).

We can now propose a measure of effectiveness C for interactive CLIR systems in which the searcher can not read the language of the retrieved documents:

$$\begin{aligned} C &= k \cdot E_i[E_j[\frac{p_r \cdot j}{r(i,j)}]] + (1-k)E_i[E_j[\frac{j + ((1-p_f)(r(i,j)-j))}{r(i,j)}]] \\ &= k \cdot p_r \cdot MAP + (1-k)(1-p_f(1-MAP)) \end{aligned}$$

where the free parameter $k \in [0,1]$ reflects the relative importance to the searcher of limiting the number of examined documents (the first term) and of limiting the translation of non-relevant documents (the second term).[5] The first term

[2] To retain a comparable form for the formula, it is also necessary to assume that the last document selected by the searcher actually happens to be relevant.

[3] This formulation does not explicitly recognize that the process may ultimately yield far too many or far too few relevant documents. If too few result, the searcher can proceed further down the list, commissioning more translations. If too many result, the searcher can adopt a more conservative strategy next time.

[4] An alternative process would be to begin at the top of the list and commission a fluent human translation of one document at a time, only proceeding to another document after examining the previous one.

[5] The linear combination oversimplifies the situation somewhat, and is best thought of here as a presentation device rather than as an accurate model of value.

reflects a straightforward adjustment to the formula for mean average precision to incorporate p_r. In the second term, success is achieved if the document is actually relevant (j) or if the document is not relevant $(r(i,j)-j))$ and is not selected by the searcher for translation $(1-p_f)$.[6] In practice, we expect one or the other term to dominate this measure. When the machine translation that is already being produced for use in the interface will suffice for the ultimate use of any document, $k \approx 1$, so:

$$C \approx p_r \cdot MAP$$

By contrast, when human translation is needed to achieve adequate fluency for the intended use, we would expect $k \approx 0$, making the second term dominant:

$$C \approx 1 - p_f(1 - MAP)$$

In either case, it is clear that maximizing MAP is desirable. When machine translation can adequately support the intended use of the documents, the factor that captures the searcher's contribution to the retrieval process is p_r (which should be as large as possible). By contrast, when human translation is necessary, the factor that captures the searcher's contribution is p_f (which should be as small as possible). This analysis suggests three possible goals for an evaluation campaign:

MAP. This has been the traditional focus of the CLIR evaluations at TREC, NTCIR and CLEF. Improvements in MAP can benefit a broad range of applications, but with 70-85% of monolingual MAP now being routinely reported in the CLIR literature, shifting some of the focus to other factors would be appropriate.

p_r. A focus on p_r is appropriate when the cost of finding documents dominates the total cost, as would be the case when present fully automatic machine translation technology produces sufficiently fluent translations.

p_f. A focus on p_f is appropriate when the cost of obtaining a translations that are suitable for the intended use dominates the total cost, as would be the case when substantial human involvement in the translation process is required. Although it may appear that $p_f = 0$ could be achieved by simply never commissioning a translation, such a strategy would be counterproductive since no relevant documents would ever be translated. The searcher's goal in this case must therefore be to achieve an *adequate* value for p_r while minimizing p_f.

The second and third of these goals seem equally attractive, since both model realistic applications. The next section explores the design of an evaluation framework that would be sufficiently flexible to accommodate either focus.

[6] For notational simplicity, p_r and p_f have been treated as if they are independent of i and j.

4 Evaluating Document Selection

Although there has not yet been any coordinated effort to evaluate cross-language document selection, we are aware of three reported user study results that have explored aspects of the problem. In one, Oard and Resnik adopted a classification paradigm to evaluate browsing effectiveness in cross-language applications, finding that a simple gloss translation approach allowed users to outperform a Naive Bayes classifier [8]. In the second, Ogden et al., evaluated a language-independent thumbnail representation for the TREC-7 interactive track, finding that the use of thumbnail representations resulted in even better instance recall at 20 documents than was achieved using English document titles [9]. Finally, Oard, et al. described an experiment design at TREC-9 in which documents judged by the searcher as relevant were moved higher in the ranked list and documents judged as not relevant were moved lower [7]. They reported that the results of a small pilot study were inconclusive. All three of these evaluation approaches reflect the effect of p_r and p_f in a single measure, but they each exploit an existing evaluation paradigm that limits the degree of insight that can be obtained. Four questions must be considered if we are to evaluate an interactive component of a cross-language retrieval system in a way that reflects a vision of how that system might actually be used:

- What process to we wish to model?
- What independent variable(s) (causes) do we wish to consider?
- What dependent variable(s) (effects) do we wish to understand?
- How should the measure(s) of effectiveness be computed?

Two processes have been modeled in the Text Retrieval Conference (TREC) interactive track evaluations. In TREC-5, -6, -7 and -8, subjects were asked to identify different instances of a topic (e.g., different countries that import Cuban sugar). This represents a shift in focus away from topical relevance and towards what is often called “situational relevance.” In the situational relevance framework, the value of a document to a searcher depends in part on whether the searcher has already learned the information contained in that document. In the TREC interactive track, subjects were not rewarded for finding additional documents on the same aspect of a topic. The TREC-9 interactive track modeled a related process in which searchers were required to synthesize answers to questions based on the information in multiple documents.

Moving away from topical relevance makes sense in the context of monolingual retrieval because the searcher’s ability to assess the topical relevance of documents by reading them is already well understood (c.f., [15]). Such is not the case in cross-language applications, where translation quality can have a substantial impact on the searcher’s ability to assess the topical relevance. An initial focus on a process based on topical relevance can thus be both informative and economical (since the same relevance judgments used to evaluate fully automatic systems can be used).

The next two questions deal with cause and effect. The complexity of an evaluation is roughly proportional to the product of the cardinality of the independent variables, so it is desirable to limit the choice of independent variables as much as possible. In the TREC, NTCIR and CLEF evaluations of the fully automatic components of CLIR systems, the independent variable has been the retrieval system design and the dependent variable has been retrieval system effectiveness. Since we are interested in the interactive components of a cross-language retrieval system, it would be natural to hold the fully automatic components of the retrieval system design constant and vary the user interface design as the independent variable. This could be done by running the automatic component once and then using the same ranked list with alternate user interface designs. Although it might ultimately be important to also consider other dependent variables (e.g., response time), retrieval effectiveness is an appropriate initial focus. After all, it would make little sense to deploy a fast, but ineffective, retrieval system.

The final question, the computation of measure(s) of effectiveness, actually includes two subquestions:

- What measure(s) would provide the best insight into aspects of effectiveness that would me meaningful to a searcher?
- How can any confounding effects that could potentially confound the estimate of the measure(s) be minimized?

When a single-valued measure can be found that reflects task performance with adequate fidelity, such a measure is typically preferred because the effect of alternative approaches can be easily expressed as the difference in the value of that measure. Mean average precision is such a measure for ranked retrieval systems. Use of a ranked retrieval measure seems inappropriate for interactive evaluations, however, since we have modeled the searcher's goal as *selecting* (rather than *ranking*) relevant documents.

One commonly used single-valued measure for set-based retrieval systems is van Rijsbergen's F measure, which is a weighted harmonic mean of recall and precision:

$$F_\alpha = 1 - \frac{1}{\frac{\alpha}{P} + \frac{1-\alpha}{R}}$$

$$\alpha = \frac{1}{\beta^2 + 1}$$

where P is the precision (the fraction of the selected documents that are relevant), R is the recall (the fraction of the relevant documents that are selected), and β is the ratio of relative importance that the searcher ascribes to recall and precision [14]. It is often assumed that $\beta = 1$ (which results in the unweighted harmonic mean), but the value for β in an interactive CLIR evaluation should be selected based on the desired balance between on p_r and p_f that is appropriate for the process being modeled.

Another possibility would be to adopt an additive utility function similar to that used for set-based retrieval evaluation in the TREC filtering track and the

Topic Detection and Tracking (TDT) evaluation:

$$C_{a,b} = N_r + a \cdot N_f + b \cdot N_m$$

where N_r is the number of relevant documents that are selected by the user, N_f is the number of false alarms (non-relevant documents that are incorrectly selected by the user), N_m is the number of misses (relevant documents that are incorrectly rejected by the user), and a and b are weights that reflect the costs of misses and and false alarms relative to correct selections.

Regardless of which measure is chosen, several factors must be considered in any study design:

- A system effect, which is what we seek to measure.
- A topic effect in which some topics may be "easier" than others. This could result, for example, from the close association of an unambiguous term (a proper name, perhaps) with one topic, while another might only be found using combinations of terms that each have several possible translations.
- A topic-system interaction, in which the effect of a topic compared to some other topic varies depending on the system. This could result, for example, if one system was unable to translate certain terms that were important to judging the relevance of a particular topic.
- A searcher effect, in which one searcher may make relevance judgments more conservatively than another.
- A searcher-topic interaction, in which the effect of a searcher compared to some other searcher varies depending on the topic. This could result, for example, from a searcher having expert knowledge on one some topic that other searchers must judge based on a less detailed understanding.
- A searcher-system interaction, in which the effect of a searcher compared to some other searcher varies depending on the system. This could result, for example, from one searcher having better language skills, which might be more important when using one system than another.
- A searcher-topic-system interaction.

In the CLEF evaluation for fully automatic CLIR, the topic has been modeled as an additive effect and accommodated by taking the mean of the uninterpolated average precision over a set of (hopefully) representative topics. In the TREC interactive track, the topic and searcher have been modeled as additive effects, and accommodated using a 2×2 Latin square experiment design. Four searchers were given 20 minutes to search for documents on each of six topics in the TREC-5 and TREC-6 interactive track evaluations [10, 11]. Eight searchers were given 15 minutes to search for documents on each of eight topics in the TREC-7 interactive track evaluation [12]. Twelve searchers were given 20 minutes to search for documents on each of six topics in the TREC-8 interactive track evaluation [4]. In each case, the Latin square was replicated as many times as the number of searchers and topics allowed in order to minimize the effect of the multi-factor interactions. Cross-site comparisons proved to be uninformative, and were dropped after TREC-6 [11]. The trend towards increasing the number of searchers reflects the difficulty of discerning statistically significant differences with a limited number of searchers and topics [4]. User studies require a substantial investment—each participant in the TREC-8 interactive track was required to obtain the services of twelve human subjects with appropriate qualifications (e.g., no prior experience with either system) for about half a day each and to develop two variants of their interactive retrieval system.

5 An Interactive CLIR Track for CLEF?

The foregoing discussion suggests that it would be both interesting and practical to explore interactive CLIR at one of the major CLIR evaluations (TREC, CLEF, and/or NTCIR). In thinking through what such an evaluation might look like in the context of CLEF, the following points should be considered:

Experiment Design. The replicated Latin square design seems like a good choice because there is a wealth of experience to draw upon from TREC. Starting at a small scale, perhaps with four searchers and six topics, would help to minimize barriers to entry, an important factor in any new evaluation. Options could be provided for teams that wished to add additional searchers in groups of 4. Allowing searchers 20 minutes per topic is probably wise, since that has emerged as the standard practice in the TREC interactive track. The topic selection procedure will need to be considered carefully, since results for relatively broad and relatively narrow topics might differ.

Evaluation Measure. There would be a high payoff to retaining an initial focus on topical relevance, at least for the first evaluation, since documents found by interactive searchers could simply be added to the relevance judgment pools for the main (fully automatic) evaluation. The F_β measure might be a good choice, although further analysis would be needed to determine an appropriate value for β once the relative importance of p_r and p_f is decided, and other measures should also be explored. The instructions given to the subjects will also be an important factor in minimizing a potential additional effect from misunderstanding the task. Subjects without formal training in relevance assessment sometimes confound the concept of topical relevance (the relationship between topic and document that is the basis for evaluation in CLEF) with the concept of situational relevance (a relationship between a searcher's purpose and a document that captures the searcher's assessment of the suitability of the document for that [possibly unstated] purpose). Providing clear instructions and adequate time for training will be essential if relevance assessments are to be obtained from subjects that are comparable to the ground truth relevance judgments produced by the CLEF assessors.

Document Language. It would be desirable to agree on a common document collection because it is well known that the performance of retrieval systems varies markedly across collections. That may be impractical in a place as linguistically diverse as Europe, however, since the choice of any single document language would make it difficult for teams from countries where that language is widely spoken to find cross-language searchers. For the first interactive cross-language evaluation, it might therefore make more sense to allow the use of documents in whichever language(s) would be appropriate for the searchers and for the translation resources that can be obtained.

Retrieval System. Interactive cross-language retrieval evaluations should focus on the interactive components of the system, so to the extent possible the fully automatic components should be held constant. If the participants agree to focus on interactive document selection, the use of a common ranked list with different interfaces would seem to be appropriate. Providing a standard ranked list of documents for each topic would help reduce barriers to entry by making it possible for a team to focus exclusively on user interface issues if that is their desire. Since cross-site comparisons were found to be uninformative in the TREC interactive track, it is probably not necessary to require the use of these standard ranked lists by every team.

Two non-technical factors will also be important to the success of an interactive cross-language retrieval track within a broader evaluation campaign. The first, an obvious one, is that coordinating the track will require some effort. A number of experiment design issues must be decided and communicated, results assembled, reports written, etc. The second, perhaps even more important, is that the track would benefit tremendously from the participation of one or more teams that already have experience in both the TREC interactive track and at least one cross-language retrieval evaluation. Several teams with this sort of experience exist, including Sheffield University in the U.K., the IBM Thomas J. Watson Research Center, New Mexico State University, the University of California at Berkeley and the University of Massachusetts at Amherst in the USA, and the Royal Melbourne Institute of Technology in Australia. With this depth of experience, the critical mass needed to jump start the evaluation process may indeed be available.

6 Forming a Research Community

CLEF is an example of what is known as an evaluation-driven research paradigm, in which participants agree on a common problem, a common model of that problem, and a common set of performance measures. Although evaluation-driven research paradigms risk the sort of local optimization that can result from choice of a single perspective, a key strength of the approach is that it can foster rapid progress by bringing together researchers that might not otherwise have occasion to collaborate, to work in a common framework on a common problem. It is thus natural to ask what about the nature of the research community that would potentially participate in an interactive CLIR evaluation. One measure of the interest in the field is that a workshop on this topic at the University of Maryland attracted eighteen participants from nine organizations and included five demonstrations of working prototype systems [1]. Another promising factor is the existance of three complementary literatures that offer potential sources of additional insights into how the cross-language document selection task might be supported: machine translation, abstracting/text summarization, and human-computer interaction.

Machine translation has an extensive research heritage, although evaluation of translation quality in a general context has proven to be a difficult problem. Recently, Taylor and White inventoried the tasks that intelligence analysts perform using translated materials and found two (discarding irrelevant documents and finding documents of interest) that correspond exactly with cross-language document selection [13]. Their ultimate goal is to identify measurable characteristics of translated documents that result in improved task performance. If that line of inquiry proves productive, the results could help to inform the design of the machine translation component of document selection interfaces.

The second complementary literature is actually a pair of literatures, alternately known as abstracting (a term most closely aligned with the bibliographic services industry) and text summarization (a term most closely aligned with research on computational linguistics). Bibliographic services that process documents in many languages often produce abstracts in English, regardless of the document language. Extensive standards already exist for the preparation of abstracts for certain types of documents (e.g., Z39.14 for reports of experimental work and descriptive or discursive studies [6]), and there may be knowledge in those standards that could easily be brought to bear on the parts of the cross-language document selection interface that involve summarization. There is also some interest in the text summarization community in cross-language text summarization, and progress on that problem might find direct application in CLIR applications. One caveat in both cases is that, as with translation, the quality of a summary can only be evaluated with some purpose in mind. Document selection requires what is known in abstracting as an “indicative abstract.” Research on “informative” or “descriptive” abstracts may not transfer as directly.

Finally, the obvious third complementary literature is human-computer interaction. Several techniques are known for facilitating document selection in monolingual applications. For example, the “keyword in context” technique is commonly used in document summaries provided by Web search engines—highlighting query terms and showing them in the context of their surrounding terms. Another example is the “show best passage” feature that some text retrieval systems (e.g., Inquery) provide. Extending ideas like these to work across languages is an obvious starting point. Along the way, new ideas may come to light. For example, Davis and Ogden allowed searchers to drill down during cross-language document selection by clicking on a possibly mistranslated word to see a list of alternative translations [2].

Drawing these diverse research communities together with the existing CLIR community will be a challenge, but there is good reason to believe that each would find an interactive CLIR evaluation to be an attractive venue. The design of tractable evaluation paradigms has been a key challenge for both machine translation and text summarization, so a well designed evaluation framework would naturally attract interest from those communities. Human-computer interaction research is an enabling technology rather than an end-user application, so that community would likely find the articulation of an important problem that is clearly dependent on user interaction to be of interest. As we have seen in the CLIR and TREC interactive track evaluations, the majority of the participants in any interactive CLIR evaluation will likely self-identify as information retrieval researchers. But experience has shown that the boundaries become fuzzier over time, with significant cross-citation between complementary literatures, as the community adapts to new challenges by integrating new techniques. This community-building effect is perhaps one of the most important legacies of any evaluation campaign.

7 Conclusion

Reviewing results from the TREC interactive track, Hersh and Over noted that "users showed little difference across systems, many of which contained features shown to be effective in non-interactive experiments in the past" [4]. Pursuing this insight, Hersh et al. found that an 81% relative improvement in mean average precision resulted in only a small (18%) and not statistically significant improvement in instance recall [5]. If this were also true of CLIR, perhaps we should stop working on the problem now. The best CLIR systems already report mean average precision values above 75% of that achieved by their monolingual counterparts, so there appears to be little room for further improvement in the fully automated components of the system. But the results achieved by Hersh et al. most likely depend at least in part on the searcher's ability to read the documents that are presented by the retrieval system, and it is easy to imaging CLIR applications in which that would not be possible without some form of automated translation. If we are to make rational decisions about where to invest our research effort, we must begin to understand CLIR as an interactive process. Beginning with a focus on the cross-language document selection process seems to be appropriate, both for the insight that it can offer and for the tractability of the evaluation.

We somewhat euphemistically refer to our globally interconnected information infrastructure as the World-Wide Web. At present, however, it is far less than that. For someone who only reads English, it is presently the English-Wide Web. A reader of only Chinese sees only the Chinese-Wide Web. We are still faced with two problems that have been with us since the Tower of Babel: how to find the documents that we need, and how to use the documents that we find. The global series of CLIR evaluations—TREC, NTCIR and CLEF—have started us on the path of answering the first question. It is time to take the second step along that path, and begin to ask how searchers and machines can work together to find documents in languages that the searcher cannot read better than machines can alone.

Acknowledgments

The author is grateful to the participants in the June 2000 workshop on interactive cross-language retrieval, and especially to Bill Ogden, Gina Levow and Jianqiang Wang, for stimulating discussions on this subject, and to Rebecca Hwa, Paul Over and Bill Hersh for their helpful comments on earlier versions of this paper. This work was supported in part by DARPA contract N6600197C8540 and DARPA cooperative agreement N660010028910.

References

1. Workshop on interactive searching in foreign-language collections. http://www.clis.umd.edu/conferences/hcil00/, June 2000.
2. Mark Davis and William C. Ogden. Quilt: Implementing a large-scale cross-language text retrieval system. In *Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 1997.
3. Marti A. Hearst. User interfaces and visualization. In Ricardo Baeza-Yates and Berthier Ribeiro-Neto, editors, *Modern Information Retrieval*, chapter 10. Addison Wesley, New York, 1999. http://www.sims.berkeley.edu/~hearst/irbook/chapters/chap10.html.
4. William Hersh and Paul Over. TREC-8 interactive track report. In *The Eighth Text REtrieval Conference (TREC-8)*, pages 57–64, November 1999. http://trec.nist.gov.
5. William Hersh, Andrew Turpin, Susan Price, Benjamin Chan, Dale Kraemer, Lynetta Sacherek, and Daniel Olson. Do batch and user evaluations give the same results? In *Proceedings of the 23nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 17–24, August 1998.
6. National Information Standards Organization. *Guidelines for Abstracts (ANSI/NISO Z39.14-1997).* NISO Press, 1997.
7. Douglas W. Oard, Gina-Anne Levow, and Clara I. Cabezas. TREC-9 experiments at Maryland: Interactive CLIR. In *The Ninth Text Retrieval Conference (TREC-9)*, November 2000. To appear. http://trec.nist.gov.
8. Douglas W. Oard and Philip Resnik. Support for interactive document selection in cross-language information retrieval. *Information Processing and Management*, 35(3):363–379, July 1999.
9. William Ogden, James Cowie, Mark Davis, Eugene Ludovik, Hugo Molina-Salgado, and Hyopil Shin. Getting information from documents you cannot read: An interactive cross-language text retrieval and summarization system. In *Joint ACM DL/SIGIR Workshop on Multilingual Information Discovery and Access*, August 1999. http://www.clis.umd.edu/conferences/midas.html.
10. Paul Over. TREC-5 interactive track report. In *The Fifth Text REtrieval Conference (TREC-5)*, pages 29–56, November 1996. http://trec.nist.gov.
11. Paul Over. TREC-6 interactive track report. In *The Sixth Text REtrieval Conference (TREC-6)*, pages 73–82, November 1997. http://trec.nist.gov.
12. Paul Over. TREC-7 interactive track report. In *The Seventh Text REtrieval Conference (TREC-7)*, pages 65–71, November 1998. http://trec.nist.gov.
13. Kathryn Taylor and John White. Predicting what MT is good for: User judgments and task performance. In David Farwell, Laurie Gerber, and Eduard Hovy, editors, *Third Conference of the Association for Machine Translation in the Americas*, pages 364–373. Springer, October 1998. Lecture Notes in Artificial Intelligence 1529.

14. C. J. van Rijsbergen. *Information Retrieval.* Butterworths, London, second edition, 1979.
15. W. John Wilbur. A comparison of group and individual performance among subject experts and untrained workers at the document retrieval task. *Journal of the American Society for Information Science*, 49(6):517–529, 1998.