Information Retrieval, 7, 205–228, 2004


# Interactive Cross-Language Document Selection

DOUGLAS W. OARD oard@umd.edu
*Human-Computer Interaction Laboratory, College of Information Studies and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA*

JULIO GONZALO julio@lsi.uned.es
*Departamento de Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia, E.T.S.I Industriales, Ciudad Universitaria s/n, 28040 Madrid, Spain*

MARK SANDERSON m.sanderson@sheffield.ac.uk
*Department of Information Studies, University of Sheffield, Western Bank, Sheffield, S10 2TN, UK*

FERNANDO LÓPEZ-OSTENERO flopez@lsi.uned.es
*Departamento de Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia, E.T.S.I Industriales, Ciudad Universitaria s/n, 28040 Madrid, Spain*

JIANQIANG WANG wangjq@glue.umd.edu
*College of Information Studies, University of Maryland, College Park, MD 20742, USA*

*Received December 6, 2002; Revised May 16, 2003; Accepted May 16, 2003*

**Abstract.** The problem of finding documents written in a language that the searcher cannot read is perhaps the most challenging application of cross-language information retrieval technology. In interactive applications, that task involves at least two steps: (1) the machine locates promising documents in a collection that is larger than the searcher could scan, and (2) the searcher recognizes documents relevant to their intended use from among those nominated by the machine. This article presents the results of experiments designed to explore three techniques for supporting interactive relevance assessment: (1) full machine translation, (2) rapid term-by-term translation, and (3) focused phrase translation. Machine translation was found to better support this task than term-by-term translation, and focused phrase translation further improved recall without an adverse effect on precision. The article concludes with an assessment of the strengths and weaknesses of the evaluation framework used in this study and some remarks on implications of these results for future evaluation campaigns.

**Keywords:** cross-language information retrieval, interactive information retrieval, evaluation of information retrieval systems, machine translation

## 1. Introduction

The broad penetration of the Internet in a diverse range of societies has important implications for cross-cultural communications, but language barriers remain a key obstacle to the full exploitation of this new medium. Language differences pose two important challenges: they limit the opportunities to learn about what is available, and they limit the ability of recipients to use the information that they find. In this article, we address the first of

these challenges, the task of finding information that may not be available in the searcher's preferred language.

Over the years, interactive information retrieval has proven to be a particularly useful paradigm for seeking information, one powerful feature of which is that it puts the searcher in control. Searchers exercise this control in two ways: by indicating what they are looking for (posing queries), and by examining what is found (selecting documents), iterating between those two processes as necessary.

When the query is posed in the same natural language as the documents that are sought (e.g., if both are in English), relatively simple search techniques based on vocabulary overlap are often suitable. The presence of multiple languages in the document collection introduces two additional challenges:

– It may not be practical for a searcher to formulate (and reformulate) queries in every possible document language.
– The searcher may not have the requisite language skills to read some of the documents that are suggested by the system. This might preclude recognition of some relevant documents, and it might also limit the searcher's ability to gain insights that would help to formulate more effective queries.

The first of these challenges has been studied extensively over the past decade, and several effective approaches to this problem of "Cross-Language Information Retrieval" (CLIR) are now known (c.f., Oard and Diekema 1998). Our focus in this article is therefore on the second challenge. Specifically, we have chosen to focus on the challenge of providing support for the task of recognizing topical relevance in documents that the searcher cannot read. In some cases (e.g., alerting the user to urgent new information), this might need to be a fully automatic process. In many applications, however, the effectiveness of fully automatic systems is limited by one or more of the following factors:

– The information need might initially be incompletely understood by the searcher.
– The information need might initially not be well articulated, either because the system's capabilities are underutilized or because the system's query language is insufficiently expressive.
– The ambiguity introduced by the use of natural (i.e., human) language within documents may cause the system to retrieve some documents that are not useful and/or to fail to retrieve some documents that are useful.

For this reason, automatic search technology is often embedded within interactive applications to achieve some degree of synergy between the machine's ability to rapidly cull through enormous collections using relatively simple techniques and a human searcher's ability to learn about their own information needs, to reformulate queries in ways that better express their needs and/or better match the system's capabilities, and to accurately recognize useful documents within a set of a limited size. The focus of this article is on the interactive document selection task. The searcher's task is to examine retrieved documents and select the ones that help to meet their information need. Here, searchers must recognize relevant documents in a language that they cannot read. There has been an extensive effort to develop

so-called "Machine Translation" (MT) systems to produce (hopefully) fluent and accurate translations for a number of language pairs, so it is natural to ask how well existing MT systems can support this task, as well as what we should do in cases where no MT system is available. We explored that question using two techniques, one using an online bilingual dictionary and a second using both an online dictionary and large text collections in each language.

The remainder of the article is organized as follows. The next section provides a brief overview of past work on cross-language information retrieval, machine translation, and evaluation. The common evaluation design is then explained, and the detailed design and results for each experiment introduced. The article concludes with some information about how these results have influenced the design of future interactive CLIR evaluation campaigns.

## 2. Background

In this section, we review past work on cross-language information retrieval, machine translation, and evaluation. Each of these fields developed separately, so we first briefly consider each in isolation. We then focus on the relatively few evaluations in which all three aspects have been combined to explore interactive CLIR.

### *2.1. Cross-language information retrieval*

Cross-language retrieval is a now a relatively well-studied problem. Over the past decade, research on CLIR has focused on development and evaluation of automatic approaches for ranking documents in a language different from that of the query. Present fully automatic techniques can do this almost as well as monolingual systems under similar conditions (on average, over a representative ensemble of queries, when evaluated using mean average precision as an effectiveness measure (Oard and Diekema 1998)). Ranking documents is only one step in a search process, however; some means of selecting documents from that list is also needed. Research on interactive retrieval strongly suggests that people are quite good at that task even when using ranked lists produced by systems that are considerably less effective at creating ranked lists than the current state-of-the-art (Hersh et al. 1998). It is an open question, however, whether a similar strategy would be effective with automatically produced translations of otherwise unreadable documents.

The typical way of evaluating ranked retrieval effectiveness is to obtain a collection of documents that are representative of those that would be searched in the actual application, create a set of queries that are representative of the way searchers are expected to express their interests in specific topics, somehow establish the relevance of each document to the topic represented by each query, and then compute a measure that reflects the density of relevant documents near the top of the list. In order to evaluate cross-language retrieval, the queries must be expressed in a language different from that of the documents. Because relevance judgments from native speakers are typically more reliable, a second version of each query is also typically prepared in the document language. Since queries are intended to be representative of how a searcher would express their information need,

the process of translating queries for use in the test collection must reflect the way an information need would be expressed in the target language. Test collections have been built in this way for more than ten languages through cooperative efforts at the Text Retrieval Conferences (TREC) in the USA, the NACSIS/NII Test Collection Information Retrieval (NTCIR) evaluations in Japan, and the Cross-Language Evaluation Forum (CLEF) in Europe.

### *2.2. Creation of translated surrogates*

To support manual selection in cross-language applications, a translated indicative surrogate for the document must be created. "Indicative" is used here in contrast to "informative" in keeping with the usual terminology for abstracting (Cleveland and Cleveland 2000). An indicative abstract is designed to provide the information that a reader would need to decide whether to read the document, while an informative abstract is designed to directly provide some of the information that a reader might be seeking (e.g., a summary of the conclusions in a scientific paper), thereby perhaps making it unnecessary for the user to obtain and examine the full document. The automatic construction of informative summaries is a challenging task; for our work we focus on the design of surrogates for an indicative rather than an informative purpose. Moreover, human-prepared abstracts and machine-prepared summaries typically exist at only a single scale, but document selection is an interactive task that can benefit from access to surrogates with variable degrees of compression. For example, search systems can offer either a list of brief summaries or the full text of a single document, under user control. This is a special case of the more general idea of multi-scale surrogates, in which system and user work together to achieve the optimal balance between conciseness and internal context.

Three factors affect the utility of translation technology for the document selection task: accuracy, fluency, and focus. By "accuracy" we mean the degree to which a translation reflects the intent of the original author. Both lexical selection (word choice) and presentation order can affect accuracy. By "fluency" we mean the degree to which a translation can be used quickly to achieve the intended purpose (in this case, document selection). Again, both lexical selection and presentation order can affect fluency. By focus, we mean the degree to which the reader's attention can be focused on the portions of a translated document that best support the intended task—in this case the recognition of relevant documents from among those nominated by the system. Highlighting query terms in the retrieved documents is an example of a strategy to manage focus.

One can think of translation or summarization as a cascade of three processes: (1) analysis, (2) transfer, and (3) generation. Most commercial Machine Translation (MT) systems implement this model directly, cascading syntactic analysis (parsing), rule-based and table-driven semantic transfer, and rule-based lexical selection and reordering for generation. In so-called "interlingual" systems, the analysis and generation stages are made more complex in order to obviate the need for a transfer stage. An alternative realization of the framework is used in so-called statistical MT systems: tokenization rules for analysis, translation probabilities learned from existing translation-equivalent texts for transfer, and language models learned from target-language texts for generation.

Three broad classes of techniques for automatic single-document summarization are compatible with our three-stage framework. The most widely explored technique is based on sentence-level selection, first representing the information content of each sentence (analysis), then selecting the sentences to be retained (transfer), and finally (trivially) generating the selected sentences. Symbolic techniques (similar to those used in commercial MT systems) and statistical techniques (similar to those used in statistical MT) have recently received increased attention.

Three broad approaches to evaluation of MT and automatic summarization have emerged: (1) human assessment of system output, (2) automatic evaluation using a reference set, and (3) use of the results to perform a task. With human assessment, accuracy and fluency (and, for summarization, focus) can be evaluated directly, but the evaluation effort must be repeated each time the system generates new output. Automatic assessment obviates this need through reference to one or more "gold standard" exemplars of desirable results. This can dramatically accelerate the iterative refinement of translation and summarization algorithms, but only a task-based evaluation can reveal the effects of putative improvements in fluency, accuracy and focus on human effectiveness when using an interactive retrieval system. We therefore focus next on what is known about evaluation of interactive retrieval systems.

### *2.3. Evaluating interactive retrieval*

The process by which searchers interact with information systems to find documents has been extensively studied (for an excellent overview, see Hearst 1999). There are two key points at which the searcher and the system interact: query formulation and document selection. We have chosen to focus on cross-language document selection in this article.

Two broad approaches to the study of user interaction have emerged, which might loosely be described as qualitative and quantitative. Qualitative studies are essentially abductive, seeking to generalize based on observed behavior. For example, by observing the behavior of novice Web searchers, we might learn how they use some newly developed feature of a system. If we learn that they are not using the new feature in the way that we had envisioned, that knowledge might be used to guide user training or system development efforts. Quantitative studies are, by contrast, deductive. In a quantitative study, we might observe that users perform some task of our choice significantly more quickly when using a newly developed feature of our system, thereby concluding that the feature meets our design objectives. In practice, most user studies include both qualitative and quantitative aspects, but practical considerations make it necessary to focus principally on one or the other when designing the study. For this article, we have chosen to focus on quantitative user studies.

Early interactive retrieval experiments were typically conducted using locally created collections with idiosyncratic variations in the experiment design. Consequently, it proved difficult to compare the results of experiments performed at different sites. The first major effort to overcome this limitation was the creation of an interactive track at the Text Retrieval Conference (TREC) in 1994. In the first three years of the track, alternative experiment designs were explored. The lack of a common reference continued to hinder cross-site

comparisons, however (Lagergren and Over 1998). In 1996 and 1997, participating teams tried using a common baseline system. It turned out that reliable comparisons were impeded by a failure to obtain statistical significance in the observed differences. The TREC interactive track continued until 2002.

Before the work reported in this article, interactive cross-language retrieval had not been the focus of any similar cooperative evaluation campaign. Indeed, the vast majority of CLIR research has focused on the automatic components of a system. Some results had, however, been reported by individual research teams:

- Resnik appears to be the first to have conducted usability tests on a task related to cross-language document selection, asking users to identify the topic of a foreign language text (Resnik 1997, Oard and Resnik 1999). He presented users with automatically produced word-by-word English translations of brief Japanese documents (directory entries) and asked them to group the documents by subject. He found that his subjects were able to categorize the translations more consistently than an automatic classifier, but less consistently than a comparable set of users were able to do when using more fluent human-prepared translations. Taylor and White later suggested (though did not test) using full machine translation for a similar task (Taylor and White 1998, White and Taylor 1998).
- The European TRANSLIB project was among the first to deploy a working CLIR system for a real application (in this case, a library catalog) (Michos et al. 1999). Questionnaires were used as a basis for qualitative evaluation. Experienced searchers reported finding query translation to be useful. It turned out, however, that people made little use of the title translation capabilities in TRANSLIB because they tended to use the system only to find documents in languages that they could read.
- The European MULINEX project also used questionnaires for qualitative evaluation of a Web-based CLIR system in which translations of automatically produced summaries were provided using the Systran MT system (Capstick et al. 1999). About half of the searchers found the query translation capabilities to be completely satisfactory, and the use of translated summaries exhibited an inverse relationship to self-reported reading skills in the document language.
- Ogden and Davis appear to have been the first to perform quantitative user studies of cross-language document selection (Ogden et al. 1999, Ogden and Davis 2000) examining Systran translations of German documents retrieved by an automatic system over 22 topics. They found that a single searcher, with no self-reported German reading skills, could identify relevant documents with an average of 99% precision and 86% recall. Judgments were measured in comparison to judgments provided by TREC relevance assessors. This is well within the normal range of inter-assessor agreement, suggesting that present MT technology may be adequate for such a task. They also ran a monolingual experiment comparing examination of titles with the use of a language-independent document thumbnail visualization in which a small sketch was presented with color-coded highlighting to indicate the locations where query terms were found. They found that users were remarkably adept at using thumbnail visualizations, assessing over twice as many documents in a fixed time with no significant loss in precision when compared to examining titles.

– Suzuki et al. performed the largest quantitative study of interactive document selection to date (Suzuki et al. 2001). Adopting a between-subjects design, they first had 64 subjects judge relevance based on word-by-word translation. A second group of 60 subjects judged automatically produced translated summaries. They found that users were able to judge the relevance of documents reasonably well using word-by-word translations of the full text, and it appeared that translated summaries were less useful for this purpose. The between-subjects design precluded direct comparison between the two conditions, however, even with the relatively large number of users that tried each condition.

Although evaluation of interactive cross-language retrieval has received some attention from researchers, little consensus on evaluation methodology has yet emerged, and little is therefore known about the relative effectiveness of alternative approaches. This stands in sharp contrast to evaluation of automated retrieval system components, for which a widely agreed evaluation methodology has led to a substantial investment in test collection development, relatively easy comparison of alternative approaches, and (in the case of cross-language retrieval) a near-doubling of retrieval effectiveness in five years. We therefore set as our goal developing an evaluation methodology for affordable, repeatable and insightful evaluation of cross-language document selection. In the next section we describe that design.

## 3. Experiment design

We chose cross-language document selection as our focus, both because the prior research pointed to effective support for cross-language document selection as an important capability for searchers who lacked reading skill, and because that choice made it possible to design a one-pass task and a decision-based metric. We chose a within-subjects quantitative user study design to compare selection effectiveness with different surrogates because our question was amenable to quantitative evaluation and because a within-subjects design offers greater statistical power than a between-subjects design (although at the cost of longer sessions). This made it possible to leverage a framework for cooperative evaluation that was developed over several years at the Text Retrieval Conference's interactive track.

Participating teams choose from two tasks: Selection of French documents or selection of English documents. We chose to support more than one document language because we ran the experiment in three different countries and we wanted to be able to recruit searchers that were not familiar with the document language. Each collection included four search topics for use in the experiment, plus a fifth practice topic. For each topic, the following resources were provided:

– Topic descriptions in English, French, and Spanish consisting of title, description, and narrative fields that served as a basis for the CLIR system's query.
– A ranked list of the top 50 documents produced automatically by a CLIR system, which is more than we expected any searcher would be able to examine in the time allowed. Using a common set of frozen ranked lists enhanced the potential for cross-site comparisons.
– The original untranslated version of each document, from the CLEF-2000 collection.

*Table 1*. Selected topics, with number (broad) or position (narrow) of relevant documents in the top 50.

| | | Relevant documents | |
|---|---|---|---|
| Topic | Summary | English | French |
| 11 (broad) | New constitution for South Africa | Total of 36 | Total of 27 |
| 13 (broad) | Conference on birth control | Total of 16 | Total of 11 |
| 17 (narrow) | Bush fire near Sydney | 1 2 3 4 6 7 | 17 29 |
| 29 (narrow) | Nobel prize for economics | 28 33 | 20 23 28 |

– An English translation of each document that was produced using the Systran Professional 3.0 MT system. The few words that Systran failed to translate were retained unchanged.

For our experiment design we selected two "broad" topics that asked about some general subject that we thought would have many aspects, and two "narrow" topics that asked about some specific event. We selected those topics from among the 40 CLEF 2000 topics by culling out topics that do not fall clearly into either category or for which the relevance of a document could likely be judged simply by looking for a proper name (e.g. *Suicide of Pierre Beregovoy*). Among the remaining set, we chose topics that we felt could be judged based solely on the topic description without the need for specialized background knowledge, and for which a number of relevant documents were present in the top-50 sets for both languages. Table 1 shows our choices and the density of relevant documents for each topic.

One interesting outcome of our topic selection process is that it turned out that the narrow topics consistently had far fewer known relevant documents in the CLEF-2000 collection than the broad topics. Thus, for this collection, "narrow" roughly equated to "sparse" and "broad" roughly equated to "dense." We also chose topic 33 (*Cancer genetics*, a broad topic) for training searchers at the outset of their session. The same standard resources (top-50 lists and baseline translations) were therefore provided for topic 33 as well.

### *3.1. Search procedure*

The task assigned to each participant in an experiment was to begin at the top of a ranked list that had been produced by a cross-language retrieval system (see above) and to determine for as many documents in the list as practical in the allowed time whether that document was "relevant," "somewhat relevant," or "not relevant" to a topic described by a written topic description. The written topic description included the text from the title, description, and narrative fields of the CLEF 2000 topic description. "Unsure" and "not judged" responses were also available.

Each four-search session was designed to be completed in about three hours, including initial training, searches, questionnaires, breaks. A maximum of 20 minutes was allowed for each topic, and participants were told that "more credit will be awarded for accurately assessing relevant documents than for the number of documents that are assessed, because

*Table 2*. Presentation order. Topics 11 & 13 broad, 17 & 29 narrow.

| Participant | Block #1 | Block #2 |
|---|---|---|
| 1 | System 1: 11-17 | System 2: 13-29 |
| 2 | System 2: 11-17 | System 1: 13-29 |
| 3 | System 1: 17-11 | System 2: 29-13 |
| 4 | System 2: 17-11 | System 1: 29-13 |
| 5 | System 1: 11-17 | System 2: 29-13 |
| 6 | System 2: 11-17 | System 1: 29-13 |
| 7 | System 1: 17-11 | System 2: 13-29 |
| 8 | System 2: 17-11 | System 1: 13-29 |

in a real application you might need to pay for a high-quality translation [of] each selected document." Participants were asked to complete eight questionnaires at specific points during their session to report on their computer/searching experience and attitudes, their language skills, their prior knowledge of each topic, and their comparative assessment of the two systems that they tried.

We adopted a within-subject design in which each participant searched each topic with some system. Participants, topics and systems were distributed using a modified Latin square design in a manner similar to that used in the TREC interactive tracks. The presentation order for topics was varied systematically, with participants that saw the same topic-system combination seeing those topics in a different order. The design made it possible to control for fatigue and learning effects to some extent. An eight-participant presentation order matrix is shown in Table 2. The minimum number of participants was set at 4, in which case only the top half of the matrix would be used. Additional participants could be added in groups of 4, with the same matrix being reused as needed.

### *3.2. Evaluation*

As our principal measure of effectiveness we selected Van Rijsbergen's $F$ measure, which is a weighted harmonic mean of precision and recall:

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

where $P$ is precision and $R$ is recall (van Rijsbergen 1979).

It is common to set $\alpha = 1/(\beta^2 + 1)$, with $\beta = 2.0 \Rightarrow \alpha = 0.2$, reflecting the case in which recall is valued twice as much as precision. Similarly, $\beta = 0.5 \Rightarrow \alpha = 0.8$, reflecting the case in which precision is valued twice as much as recall.[1] For this evaluation, we chose $\beta = 0.5$, modeling the case in which missing some relevant documents would be less objectionable than finding too many documents that, after perhaps paying for professional translations, turn out not to be relevant (Oard et al. 2001). The CLEF relevance judgments

are two-state (relevant or not relevant), so we treated all judgments other than "relevant" ("somewhat relevant", "not relevant", and "unsure") as not relevant when computing $F_{\beta=0.5}$. For contrast, we also computed $F_{\beta=2.0}$ (which modeled a recall-biased searcher). We also computed a contrastive condition that we called "loose relevance" in which "somewhat relevant" documents were treated as relevant when computing $F_{\beta=0.5}$.

In the minimal 4-searcher design, two searchers ran each topic-system pair. For the 8-searcher design, four searchers ran each topic-system pair. We calculated $F_\beta$ separately for each searcher-topic-system run and computed the mean across the two (or four) searchers to compute an expected value of $F_\beta$ for each topic-system pair. We then computed the mean across the two broad topics to find the expected value of $F_\beta$ for each combination of system and topic type. To test for statistical significance, we used a linear mixed effects model (Pinheiro and Bates 2000) to distinguish between the system effect that we seek to detect and the combined searcher/topic/system effect that we wish to suppress, claiming statistical significance if an analysis of variance (ANOVA) reports $p < 0.05$. The suitability of alternative models was explored by examining residuals between the model and the fitted responses, comparing quantiles of a normal distribution with the quantiles observed in our data. In the notation of the `nlme` linear mixed-effects models library of the *R* statistical package, the model that best fit the Maryland and UNED experiments was:

$$F_{\beta=0.5} \sim System, random = \sim 1 + System + Topic \mid User$$

where $F_{\beta=0.5}$ is the outcome variable, *System* is a fixed effect, and *User* is a random effect that interacts with *System* and *Topic*.

## 4. Experiments

We conducted our experiments at three sites: Universidad Nacional de Educación a Distancia (UNED) in Spain, the University of Maryland (UMD) in the United States, and the University of Sheffield (SHEF) in the United Kingdom. In the next section, we describe experiments at the University of Maryland comparing the ability of English speakers to use term-by-term gloss translations of French documents with results obtained using the baseline Systran translations. That is followed by a section in which we describe experiments at UNED that explored the ability of Spanish speakers to use phrase translations of English documents rather than full machine translation. Finally, we report on experiments at the University of Sheffield in which English speakers searched both the English documents and English translations of the French documents.[2]

### *4.1. Gloss translation experiments*

At the University of Maryland, we are interested in the development of systems for retrieval of documents in languages for which few resources exist. We therefore chose to compare term-by-term translation (which we refer to as "gloss translation") with Systran. Machine translation into English is presently available for about 40 of the world's several thousand languages; gloss translation is an alternative approach that can easily be implemented for

any of the remaining languages for which a simple bilingual term list of paired translations is available. We chose French to simulate a resource-limited language because knowledge of French among the pool of possible participants at Maryland was more limited than knowledge of English. In order to control timing effects, gloss translation was performed in advance. We implemented a backoff strategy that first translated multiword expressions that could be found in a 35,000-term English-French term list, and then translated remaining words that could be found in the term list individually. Any remaining words were then stemmed and translated using a stemmed term list; if none of this worked, the French term was presented unchanged. Figure 1 shows the results of this process for some document titles; the same process was used for the full documents.

The hypothesis that we wished to test was that gloss translation could support effective interactive cross-language document selection. Formally, we sought to reject the null hypotheses that the $F_{\beta=0.5}$ measure achieved using the MT system is the same as that which would be achieved using the gloss translation system. We sought to minimize the effect of presentation differences by using the same user interface with both types of translation.

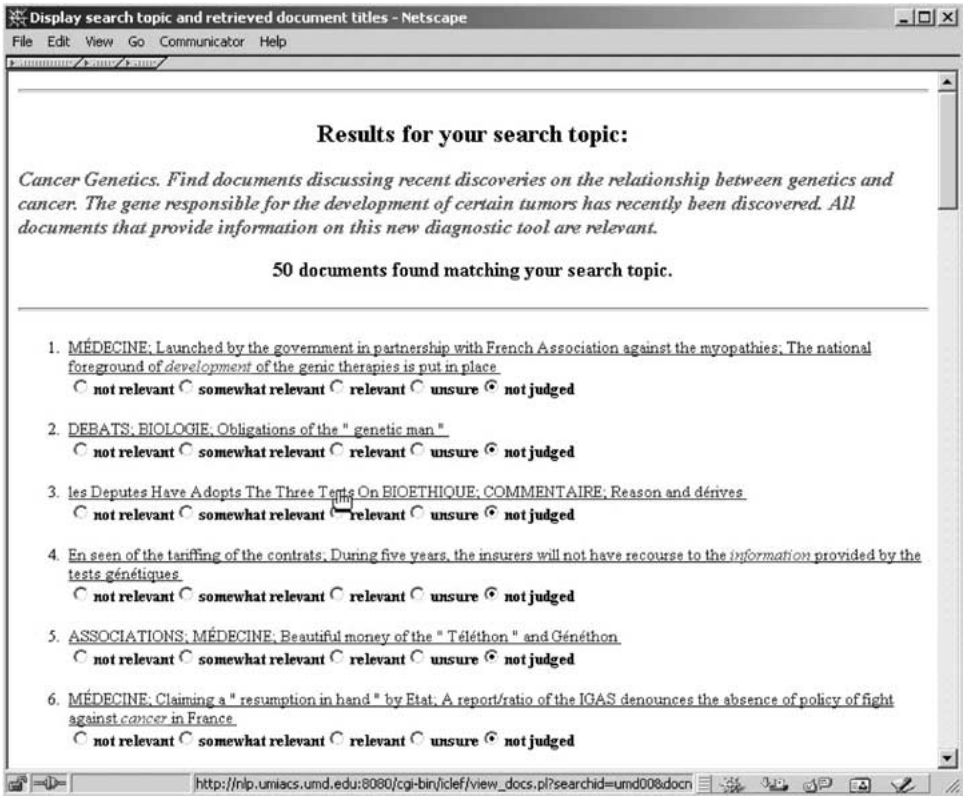


*Figure 1*. Maryland's user interface, showing the ranked list of surrogates for the Systran condition and the relevance judgment radio buttons.

*Table 3*. Maryland: $F_{\beta=0.5}$ by topic type and system.

| | Broad | | Narrow | | Average | |
|---|---|---|---|---|---|---|
| Topic searcher | MT | Gloss | MT | Gloss | MT | Gloss |
| umd01 | 0.62 | 0.28 | 1.00 | 0.78 | 0.81 | 0.53 |
| umd02 | 0.34 | 0.13 | 0.78 | 0.00 | 0.56 | 0.07 |
| umd03 | 0.13 | 0.10 | 1.00 | 0.00 | 0.52 | 0.05 |
| umd04 | 0.13 | 0.27 | 0.9 | 0.83 | 0.52 | 0.55 |
| Average | 0.31 | 0.20 | 0.92 | 0.41 | 0.61 | 0.29 |

Searchers interacted with our system using a Web browser, and their relevance judgments were recorded by a central server when a search was completed. After a small pilot study to refine our interface design and data collection methods, we conducted a total of 16 trials, with four searchers performing four searches each. None of our searchers were involved in previous interactive retrieval experiments, and all had at least five years of online searching experience. We offered each searcher a cash payment of $20.

Table 3 shows both individual and aggregate results. Three of the four searchers did better with MT than with gloss translation on broad topics, and all four did better with MT on narrow topics. An ANOVA across the 16 observations found that MT resulted in a significantly higher $F_{\beta=0.5}$ measure (at $p < 0.02$), so we can reject our null hypothesis and conclude that MT is more suitable than our implementation of gloss translation for this task.

The values of $F_{\beta=0.5}$ for narrow topics are consistently higher in Table 3 than the values for broad topics. Since most unjudged documents were for broad topics, for which an average of almost 40% of the documents were relevant, our measures penalized searchers more for failing to finish their judgments for broad than for narrow topics. If a searcher had simply marked all 50 documents for each topic as relevant, the resulting value for $F_{\beta=0.5}$ would be 0.26. All of four searchers beat that value by at least a factor of two when using the MT system, and two of the four also were able to do so when using gloss translation. The other two did quite poorly with gloss translation. From this we conclude that both MT and gloss translation can be useful, but that there is substantial variation across the population of searchers with regard to their ability to use gloss translations as a basis for document selection.

Examining the time required to make relevance judgments provides another perspective on our results. As figure 2 shows, "unsure" and "somewhat relevant" judgments took longer on average than "relevant" judgments, and "not relevant" judgments could be performed the most quickly. This was true for both topic types, and it helps to explain why narrow topics (which have few relevant documents) had fewer "not judged" cases. One possible explanation for this would be a within-topic learning effect, in which searchers learn to recognize documents in a category based on their recollection of documents that have been previously assigned to that category. A total of 398 "not relevant" judgments (the fastest category overall) were made, but only 20 "unsure" judgments (the slowest category). We

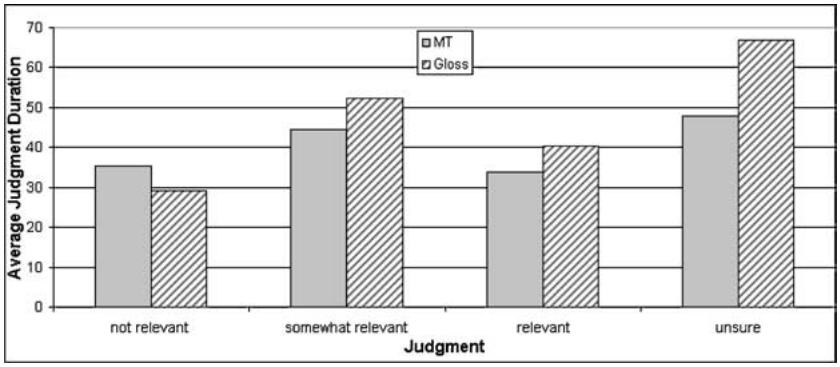


*Figure 2*. Maryland: Average time per judgment, by judgment type, for judgments with two or more observations; broad topics on the left, narrow on the right.

observed that some searchers often modified their relevance judgment for one document after examining a different document. This tends to support our inference of a within-topic learning effect, since presumably their refined judgment was informed by something that they learned about the topic by reading a document.

After each session, we solicited comments from our searchers on the two systems. All searchers reported that it was hard to comprehend meaning with gloss translations, and three of the four indicated that judging the relevance of documents using gloss translations was difficult. All four searchers felt that it was easy to make relevance judgments with the MT system, and three of the four indicated that they liked the translation quality (with no comment on this point from the fourth). Two searchers felt that the difficulty of learning to use the two systems was comparable, while the other two felt that the MT system was easier to learn. Three of the four found the MT system easier to use.

### *4.2. Phrase translation experiments*

Two translation techniques were compared at UNED: Systran translations as the reference system, and a noun-phrase translation approach based on "comparable" corpora of separately authored news stories with similar topical coverage. Our phrase translation technique used example-based techniques to avoid some of the disfluencies that are common in machine translation results, and it incorporated a natural focus mechanism (selective translation) that was further enhanced through confidence-based highlighting. The hypothesis being tested was that sufficiently accurate relevance judgments could be performed more rapidly based on phrase translation than based on full MT. Formally, we sought to reject the null hypothesis that both systems would achieve comparable levels of recall in a time-constrained search.

We used the phrase extraction software from the *UNED WTB Multilingual search engine* (Peñas et al. 2001). For each English noun phrase, we translated all non-stopwords using a bilingual dictionary. For each word in the set of translations, we considered all Spanish

phrases that contain that word. We found a total of 26,700,000 different Spanish noun phrases in the CLEF-2000 Spanish "EFE" collection of 250,000 newswire documents from 1994. Of these, we retained only the 3,600,000 phrases that appeared more than once in the collection. The set of all Spanish phrases that contained at least one translation formed a *pool of related Spanish phrases*. We then identified all phrases in this pool that contain exactly one translation for each term of the original English phrase. This subset of the pool was our *set of candidate translations*. For example, the system found:

| | Phrase | Frequency |
|---|---|---|
| abortion issue ⇒ | tema del aborto | 16 |
| | asunto del aborto | 12 |
| | asuntos como el aborto | 5 |
| | asuntos del aborto | 2 |
| | temas como el aborto | 2 |
| | asunto aborto | 2 |

If the resulting set was non-empty (as in the example above), the system selected the noun phrase in that set that occurred most often in the EFE collection as the optimal translation. Therefore "*tema del aborto*" was (correctly) chosen as translation for "*abortion issue*". If the set of candidate translations was empty, the following two steps would be taken:

1. *Subphrase translation*: The system looked for maximal sub-phrases using the algorithm described above. These were used as partial translations.
2. *Word by word contextual translation*: The remaining words were translated using phrase statistics to take context into account: from all translation candidates for a word, we chose the candidate that is included in the most phrases in the original pool of related Spanish phrases. Words for which no translation was known were retained unchanged (in the hope that they might be named entities or some other form of cognate that would be recognizable).

English phrases that were entirely subsumed by other phrases from the same document were deleted, and the remaining English phrases were displayed in the order they appear within the document; phrases with an optimal alignment were highlighted using a bold font, and phrases containing query terms were displayed in a distinctive color (bright green). Figure 3 shows an example of our phrase translation interface.

We performed three experiments with different searcher populations: for the main experiment, we recruited eight volunteers that self-reported low (or no) proficiency in the English language. For comparison, we also formed two additional eight-searcher groups that self-reported mid-level and high-level English skills, respectively. Most searchers used the system locally, but five performed the experiment from a remote location (in the presence of the same observer) using an Internet connection. This proved to be problematic, since network delays altered the interactive search experience. This effect invalidated the

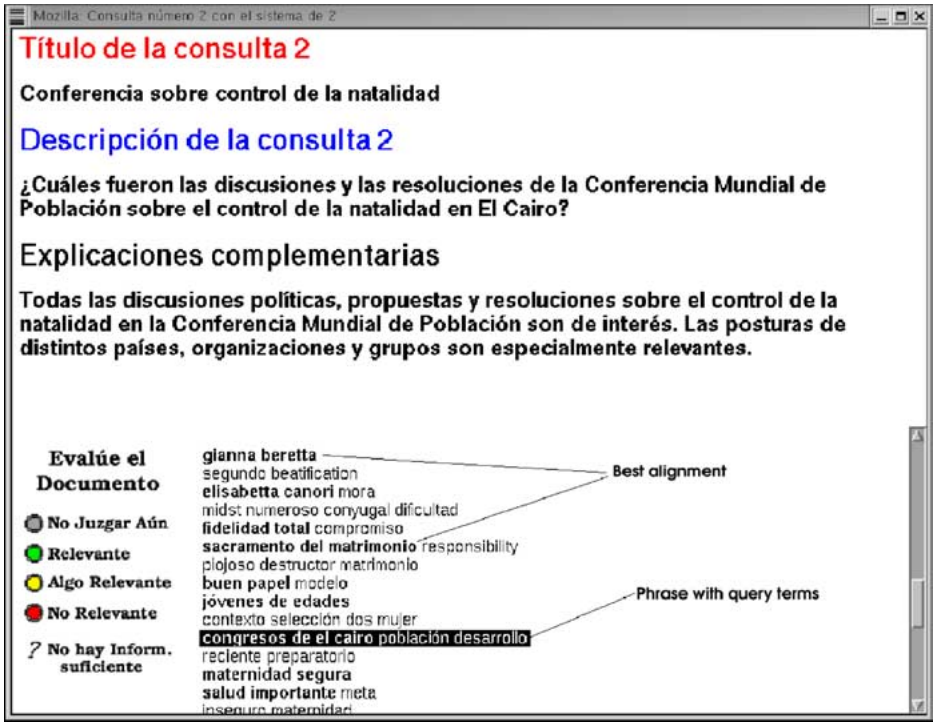


*Figure 3*. UNED search interface, phrase translation system.

mid-level group's results (with three remote searchers). The low and high proficiency groups each included only one remote searcher. The results for the main experiment (low-level group) are detailed in Table 4.

Searchers with low English skills achieved similar precision for both translation approaches, but phrase translation yielded 52% greater recall. Searchers with high English skills did somewhat better overall, but still showed a similar pattern. From this we concluded that searchers were clearly able to judge documents more quickly with little loss in accuracy when using the phrase translation interface. Remarkably, the difference between MT and phrases comes mostly from the broad topics. The most likely explanation is that relevance judgement on broad topics demands a more detailed scanning of the document contents, something that can be done faster with phrase-based summaries than with full translations.

An ANOVA on the 32 observations for the low English proficiency group revealed no significant differences in $F_{\beta=0.5}$ ($p = 0.20$) or recall ($p = 0.14$), so we cannot reject the null hypothesis. The trend across both groups seems clear, so we expect that our inability to see statistical significance results from the small amount of available data and from two searchers that exhibited unusual behavior. Searcher uned05 apparently did not understand the task, since almost no relevant documents were marked in any of the four search sessions. From questionnaire responses, it appears that searcher uned05 was actually attempting to

*Table 4*. UNED: $F_{\beta=0.5}$ by topic type and system.

| | Broad | | Narrow | | Average | |
|---|---|---|---|---|---|---|
| Topic searcher | MT | Phrase | MT | Phrase | MT | Phrase |
| uned01 | 0.09 | 0.00 | 0.48 | 0.00 | 0.29 | 0.00 |
| uned02 | 0.25 | 0.53 | 0.83 | 0.70 | 0.54 | 0.62 |
| uned03 | 0.30 | 0.38 | 0.90 | 0.27 | 0.60 | 0.33 |
| uned04 | 0.08 | 0.38 | 0.00 | 0.35 | 0.04 | 0.37 |
| uned05 | 0.00 | 0.19 | 0.00 | 0.00 | 0.00 | 0.10 |
| uned06 | 0.00 | 0.42 | 0.18 | 0.58 | 0.09 | 0.50 |
| uned07 | 0.38 | 0.00 | 0.83 | 0.35 | 0.61 | 0.18 |
| uned08 | 0.03 | 0.82 | 0.27 | 0.71 | 0.15 | 0.77 |
| Average | 0.14 | 0.34 | 0.44 | 0.37 | 0.29 | 0.36 |

judge translation quality rather than relevance. The second potentially problematic searcher was uned01, who was the only member of the low English group to perform the task remotely.

Unlike the Maryland study, most UNED searchers reported little experience with search engines. Most reported a preference for phrase translation, arguing that the information was more concise and thus decisions could be made faster, although several searchers also remarked that phrase translation demanded more interpretation from the user. The MT system was perceived as giving more detailed information, although the density of that information sometimes made the relevance judgment process difficult. These impressions are consistent with the quantitative results that we obtained, and they tend to confirm our hypothesis about the utility of noun-phrase translation as a basis for assessing topical relevance.

In summary, although the quantitative results did not reveal statistically significant differences, the combination of quantitative evidence and searcher impressions indicates that summarized translations, and in particular noun-phrase translation into the searcher's language, could be a useful feature for assessing broad topics, even when full machine translation is available. The computational cost of producing noun-phrase translations is significantly lower than that of full MT; our current implementation is at least one order of magnitude faster than Systran translation (although some of that speed advantage results from caching all possible two-word and three-word noun phrases).

### *4.3. Native speaker experiments*

The experiments at Sheffield employed only monolingual English searchers. Eight university students for whom English was their native language participated in the Sheffield experiments. Each saw untranslated English documents for two topics and Systran translations of French documents for the other two topics. Although none of the searchers regarded

themselves as French speakers, some had taken French at school in their early teenage years. We paid each searcher £20 (∼$30).

Unlike the other experiments described in this article, we cannot make meaningful within-site comparisons in this case because the difference in relevance judgements across the two sets of documents could be attributed to a broad range of factors, including the quality of the Systran translations, differences in the number of relevant documents (240 in English, 170 in French), stylistic differences between the two sources, the extent of prior cultural knowledge among the searchers, the fact that CLEF judgments are performed by different assessors for each language, and the fact that different retrieval systems were used to produce the ranked lists in each language. Our experiments therefore had two objectives that did not involve hypothesis testing:

– To characterize the degree of agreement between results obtained for the same task at diferent sites. Formally, we sought to determine the difference in $F_{\beta=0.5}$ achieved by searchers examining Systran translations of French documents at Maryland and Sheffield.
– To characterize the difference between interactive experiments and the CLEF relevance assessment process. Formally, we sought to determine whether the overlap between CLEF relevance assessors and searchers interactively examing (monolingual) English retrieval results was within the range normally seen between relevance assessors.

The results of the Sheffield experiments are shown in Table 5. In all of our experiments, we compared relevance judgements made by searchers reading translated documents to judgments made by assessors who were reading untranslated documents. Although our aim was to assess the extent to which some type of translation impaired the searcher's ability to judge topical relevance, any such measurement necessarily confounds a number of factors. One important factor is that all relevance assessments are subjective, depending on the user's interpretation of the topic statement and documents. The overlap in the sets of relevant documents judged by different assessors is commonly used as a measure of

*Table 5*. Sheffield: $F_{\beta=0.5}$ by topic type and system.

| | Broad | | Narrow | | Average | |
|---|---|---|---|---|---|---|
| Topic searcher | MT | Mono | MT | Mono | MT | Mono |
| shef01 | 0.73 | 0.30 | 0.83 | 0.28 | 0.74 | 0.30 |
| shef02 | 0.45 | 0.63 | 0.91 | 0.83 | 0.53 | 0.67 |
| shef03 | 0.56 | 0.36 | 0.00 | 0.00 | 0.47 | 0.35 |
| shef04 | 0.64 | 0.31 | 0.91 | 0.71 | 0.68 | 0.40 |
| shef05 | 0.75 | 0.37 | 1.00 | 0.36 | 0.77 | 0.37 |
| shef06 | 0.55 | 0.44 | 0.79 | 0.56 | 0.61 | 0.46 |
| shef07 | 0.42 | 0.63 | 0.00 | 0.56 | 0.41 | 0.61 |
| shef08 | 0.19 | 0.38 | 0.91 | 0.71 | 0.39 | 0.45 |
| Average | 0.54 | 0.43 | 0.67 | 0.50 | 0.58 | 0.45 |

agreement, defining overlap as the size of the set intersection divided by the size of the set union. If assessors were in perfect agreement, overlap would be 1; with no agreement at all, overlap would be 0. Voorhees found pairwise overlap values that ranged between 0.42 and 0.49 for TREC assessors, but that the preference order of two retrieval systems over an ensemble of 50 topics was rarely reversed when switching assessors (Voorhees 1998). Sanderson helped to explain this result, finding that assessors exhibit more agreement on the relevance of documents ranked more highly by ranked retrieval systems, exactly the documents that dominate typical effectiveness measures such as mean average precision (Sanderson 1998).

Our studies were limited to the top 50 documents, so if all other factors were equal, we would have expected to observe higher overlap than Voorhees. We found, however, that the overlap between the monolingual Sheffield searchers and the CLEF assessors ranged from 0.39 to 0.47 (when "somewhat relevant" judgements were treated as "relevant"). These relatively low values likely result from several effects: (1) our searchers had to make their judgments in a sharply limited period (if they had actually tried to judge every document, they would have had an average of just 24 seconds for each one); (2) CLEF assessors must judge every document as relevant or not relevant, while our searchers could also choose "somewhat relevant" or "unsure," or they could leave the document unjudged; (3) our searchers were given instructions that were intended to bias them in favor of precision—relevance assessment for CLEF, by contrast, placed a premium on careful consideration of every document in the assessment pool; (4) CLEF assessors can discuss difficult judgments with other assessors, thereby reflecting some degree of community consensus in those cases—our searchers produced only personal opinions; and (5) CLEF assessors evaluate documents in an arbitrary order, while our searchers had additional information available (in particular, the order of the documents in the ranked list).

Figure 4 illustrates the French results for each condition that was run, averaged across all searchers and both topic types. Figure 5 provides a similar depiction for English. A naive searcher that marked every document as relevant would achieve a precision of 0.30 for English or 0.22 for French. All of our results exceed those values by a substantial margin, indicating that searchers are clearly able to make good use of any of the surrogates that we have produced.

The relatively small differences between the Sheffield and Maryland MT results on French documents (which correspond to $F_{\beta=0.5}$ values of 0.59 and 0.61) provides some insight into the effect of implementation details such as user interface design. The observed difference between MT and gloss translation and between MT and phrase translation are far larger, suggesting that they reflect real differences in the relative utility of those three types of surrogates. Moreover, it seems reasonable to conclude that cross-site comparisons using a common experimental design are indeed feasible, but only when run on the same document collection.

## 5. Drawing the results together

In this section, we draw together results from all three participating teams to examine our evaluation methodology.

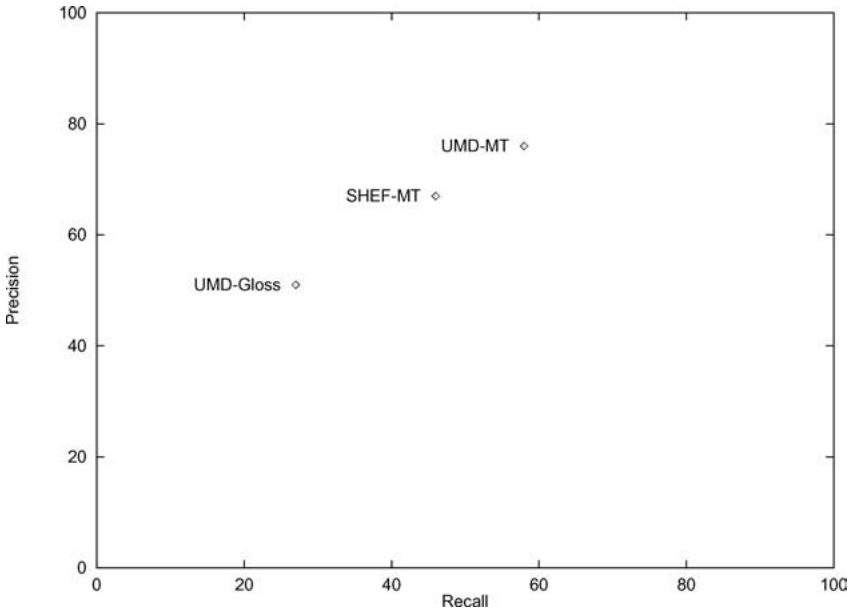


*Figure 4*. Overview of French results.

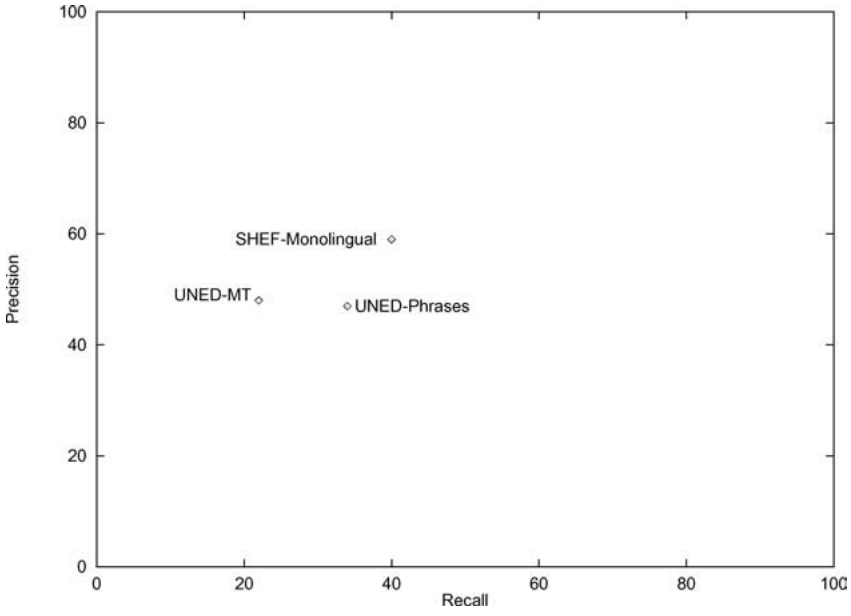


*Figure 5*. Overview of English results.

### 5.1. *Recall-oriented measures*

We originally chose a precision oriented measure ($F_{\beta=0.5}$) because we expected that coverage gaps in the available translation resources would preclude achieving high recall. We therefore instructed our searchers to seek high precision. Of course, we cannot go back and change the instructions to the searchers, but we can reanalyze our results with an alternative measure ($F_{\beta=2.0}$) that emphasizes recall in order to explore the behavior of that measure. Table 6 illustrates this idea for the overall summary data contained in figures 4 and 5. As can be seen, the preference order between conditions is preserved.

Table 7 illustrates a similar comparison at a finer level of granularity (with the last line copied from Table 4). Again, a similar preference order is evident for broad topics, although in the case of narrow topics (for which the differences are smaller) the preference order is reversed. From these results, we conclude that our results are not strongly dependent on the bias in our $F_\beta$ measure.

*Table 6*. Comparing precision-oriented and recall-oriented measures.

| System | $P$ | $R$ | $F_{\beta=0.5}$ | $F_{\beta=2.0}$ |
|---|---|---|---|---|
| | English documents | | | |
| SHEF-Monolingual | .59 | .40 | .45 | .39 |
| UNED-Phrases | .47 | .34 | .35 | .32 |
| UNED-MT | .48 | .22 | .28 | .21 |
| | French documents | | | |
| UMD-MT | .76 | .58 | .61 | .57 |
| SHEF-MT | .67 | .46 | .59 | .48 |
| UMD-Gloss | .51 | .27 | .29 | .26 |

*Table 7*. UNED: $F_{\beta=2.0}$ by topic type and system.

| | Broad | | Narrow | | Average | |
|---|---|---|---|---|---|---|
| Topic searcher | MT | Phrase | MT | Phrase | MT | Phrase |
| uned01 | 0.02 | 0.00 | 0.19 | 0.00 | 0.11 | 0.00 |
| uned02 | 0.40 | 0.22 | 0.55 | 0.90 | 0.48 | 0.56 |
| uned03 | 0.09 | 0.69 | 0.70 | 0.41 | 0.40 | 0.55 |
| uned04 | 0.06 | 0.13 | 0.00 | 0.45 | 0.03 | 0.29 |
| uned05 | 0.00 | 0.18 | 0.00 | 0.00 | 0.00 | 0.09 |
| uned06 | 0.00 | 0.15 | 0.35 | 0.63 | 0.17 | 0.39 |
| uned07 | 0.13 | 0.00 | 0.55 | 0.45 | 0.34 | 0.23 |
| uned08 | 0.05 | 0.60 | 0.41 | 0.38 | 0.23 | 0.49 |
| Average | 0.16 | 0.25 | 0.34 | 0.40 | 0.22 | 0.33 |
| ($F_{\beta=0.5}$) | 0.14 | 0.34 | 0.44 | 0.37 | 0.29 | 0.36 |

*Table 8*. Comparing strict and loose relevance judgements, $F_{\beta=0.5}$.

| | Maryland | | UNED | | Sheffield | |
|---|---|---|---|---|---|---|
| | MT | Gloss | Phrases | MT | MT | English |
| Strict | 0.61 | 0.29 | 0.36 | 0.29 | 0.60 | 0.46 |
| Loose | 0.67 | 0.42 | 0.38 | 0.40 | 0.59 | 0.52 |

### *5.2. Treatment of "somewhat relevant" documents*

We originally chose a "strict" definition of relevance in which documents marked "somewhat relevant" were treated as not relevant when computing precision and recall. We chose this approach because it modeled the ultimate use that we envisioned for the selected documents (submission for professional translation). CLEF relevance assessors are, however, instructed to treat documents with any substantial discussion of a topic as relevant. In order to assess the effect of this difference, we reanalyzed our overall results with a "loose" definition of relevance in which "somewhat relevant" documents were treated as relevant. Table 8 shows an effect of this change: for each site, the better of their two results (shown on the left) showed relatively little change, but the value of $F_{\beta=0.5}$ increased markedly for the lower-scoring condition. Moreover, with one exception (when the values are almost identical), values computed with loose relevance judgments are higher. We have seen similar results from a more detailed analysis as well (Wang and Oard 2001). From this consistent evidence, we conclude that our searchers most likely used the "somewhat relevant" category in preference to "unsure" in cases where they observed some evidence of relevance but were unable to positively establish the relevance of the document.

## 6. Conclusions

Cooperative evaluations such as iCLEF offer three potential benefits, all of which are present in this case:

*Insight*. All three of the surrogates that we tried (term-by-term gloss translation, full machine translation, and phrase translation) proved to be useful, and we found a clear preference ordering among them.

*Consensus*. We agreed on a common evaluation framework that others can use to replicate our work, or to explore additional contrastive conditions. All of the materials that we used are available to any team participating in the Cross-Language Evaluation Forum.

*Community*. The iCLEF evaluations continued in 2002, with participation from five research teams in four countries, and an another evaluation is now planned for 2003. Moreover, the track has emerged as one of the principal evaluation venues for the Clarity project, in which seven institutions in four countries are cooperating to develop and evaluate interactive cross-language information retrieval technology.

Our experience in this first year of iCLEF has shaped our thinking for subsequent evaluations in the following ways:

- Relevance judgment is an important task, but actual interactive information seeking processes are often considerably more complex. For iCLEF 2002, we added an interactive query formulation task, with allowances for query reformulation and relevance feedback. For the teams that elected that more complex task, $F_{\beta=0.5}$ served as an outcome measure, with other measures being developed to provide insight into the iterative query reformulation process. The relevance judgment task was retained as an alternative to full interactive search, both to facilitate more focused studies and to minimize the entry costs for first-year participants.
- Our results were consistent with a hypothesis that some searchers used "somewhat relevant" to report uncertainty rather than using the "unsure" category that we had provided for that purpose. We therefore separated the reporting of the degree of relevance and the confidence in that judgment for iCLEF 2002 and 2003.
- In the Maryland experiments, we noted that the $F_{\beta=0.5}$ evaluation measure seemed to behave somewhat differently with broad and narrow topics, with a greater tendency toward extreme values (zero and one) for narrow topics, and a stronger central tendency (towards the mean) for broad topics. The relatively small number of relevant documents for narrow topics can introduce substantial quantization errors, which might explain that effect. In order to limit the sources of unintended variation in future experiments, we decided to include only broad topics in 2002 and 2003.

Our experiment design could also be extended in a number of interesting ways. For example, differences between monolingual and cross-language retrieval effectiveness might be also evaluated on the same document collection using either bilingual searchers or a between-subjects experiment design. In either case, a larger number of searchers would likely be required to obtain similar confidence in observed differences. Creation of multi-valued relevance judgments (e.g., yes/partial/no tri-state judgments, or aspect-segregated relevance) might also make it possible to model realistic use cases for interactive cross-language retrieval with higher fidelity.

Regardless of the refinements that we introduce, experimental evaluation of interactive retrieval systems is an expensive undertaking. It is, however, a highly leveraged investment. Over the past decade, we have developed a broad array of component technologies for cross-language retrieval, machine translation, and automatic generation of summaries. By harvesting the techniques that perform best in relatively inexpensive automated evaluations, we can use interactive evaluations to explore potential synergies between interconnected components and to help us understand the limitations of our automated evaluation techniques. The experiments described in this article represent a first step in that direction.

## Acknowledgments

The authors are grateful to Robert Allen, Zoë Bathie, Clara Cabezas, Bill Hersh, Gina Levow, Paul McNamee, Fermín Moscoso del Prado, Paul Over, Carol Peters and Daniela Petrelli

for their help. This work has been supported in part by DARPA cooperative agreement N660010028910 (TIDES), EU projects IST-2000-25310 (Clarity) and IST-2000-26061 (MIND), and the Spanish government project TIC-2000-0335-C03-01 (Hermes).

## Notes

1. Formally, $\partial F_\beta / \partial P = \partial F_\beta / \partial R$ for $R = \beta P$.
2. Additional details on the experiments run at each site can be found in López-Ostenero et al. (2001), Sanderson and Bathie (2001) and Wang and Oard (2001).

## References

Capstick J, Diagne AK, Erbach G, Uzkoreit H, Leisenberg A and Leisenberg M (1999) A system for supporting cross-lingual information retrieval. Information Processing and Management, 36(2):275–289.

Cleveland DB and Cleveland AD (2000) Introduction to Indexing and Abstracting, 3rd edn. Libraries Unlimited, Englewood, CO.

Hearst MA (1999) User interfaces and visualization. In: Baeza-Yates R and Ribeiro-Neto B, Eds. Modern Information Retrieval. Addison Wesley, New York, Chapt. 10.

Hersh W, Turpin A, Price S, Chan B, Kraemer D, Sacherek L and Olson D (1998) 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, pp. 17–24.

Lagergren E and Over P (1998) Comparing interactive information retrieval systems across sites: The TREC-6 interactive track matrix experiment. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

López-Ostenero F, Gonzalo J, Peñas A and Verdejo F (2001) Noun phrase translations for cross-language document selection. In: Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF-2001. Revised papers. Springer-Verlag, LNCS 2406.

Michos S, Stamatatos E and Fakotakis N (1999) Supporting multilinguality in library automation systems using AI tools. Applied Artificial Intelligence.

Oard DW and Diekema AR (1998) Cross-language information retrieval. In: Annual Review of Information Science and Technology, Vol. 33, American Society for Information Science.

Oard DW, Levow G-A and Cabezas CI (2001) CLEF experiments at Maryland: Statistical stemming and backoff translation. In: Peters C, Ed. Proceedings of the First Cross-Language Evaluation Forum. Cross-Language Information Retrieval and Evaluation. Springer-Verlag, LNCS 2069.

Oard DW and Resnik P (1999) Support for interactive document selection in cross-language information retrieval. Information Processing and Management, 35(3):363–379.

Ogden W, Cowie J, Davis M, Ludovik E, Molina-Salgado H and Shin H (1999) 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.

Ogden WC and Davis MW (2000) Improving cross-language text retrieval with human interactions. In: Proceedings of the 33rd Hawaii International Conference on System Sciences.

Peñas A, Gonzalo J and Verdejo F (2001) Cross-language information access through phrase browsing. In: Applications of Natural Language to Information Systems, pp. 121–130.

Pinheiro J and Bates D (2000) Mixed-Effects Models in S and S-PLUS. Springer.

Resnik P (1997) Evaluating multilingual gisting of Web pages. In: AAAI Symposium on Cross-Language Text and Speech Retrieval.

Sanderson M (1998) Accurate user-directed summarization from existing tools. In: Proceedings of the 7th International Conference on Information and Knowledge Management.

Sanderson M and Bathie Z (2001) iCLEF at Sheffield. In: Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF-2001. Revised papers. Springer-Verlag, LNCS 2406.

Suzuki M, Inoue N and Hashimoto K (2001) A method for supporting document selection in cross-language information retrieval and its evaluation. Computers and the Humanities, 35(4):421–438.

Taylor K and White J (1998) Predicting what MT is good for: User judgments and task performance. In: Farwell D, Gerber L and Hovy E, Eds. Third Conference of the Association for Machine Translation in the Americas, Springer. Lecture Notes in Artificial Intelligence 1529, pp. 364–373.

van Rijsbergen CJ (1979) Information Retrieval, 2nd edn. Butterworths, London.

Voorhees E (1998) Variations in relevance judgments and the measurement of retrieval effectiveness. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Wang J and Oard DW (2001) iCLEF 2001 at Maryland: Comparting term-for-term and gloss translations. In: Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF-2001. Revised papers. Springer-Verlag, LNCS 2406.

White JS and Taylor KB (1998) A task-oriented evaluation metric for machine translation. In: First International Conference on Language Resources and Evaluation, pp. 21–25.