Studying the Use of Interactive Multilingual Information Retrieval

Daqing He[1], Douglas W. Oard[2], and Lynne Plettenberg[2]

[1]School of Information Sciences, University of Pittsburgh, 135 North Bellefield Avenue, Pittsburgh, PA 15260
daqing@mail.sis.pitt.edu

[2]College of Information Studies, University of Maryland, College Park, MD 20742
{oard, lpletten}@umd.edu

ABSTRACT

We often talk as if information retrieval systems were machines, but in reality the "systems" that we use to retrieve information are synergistic combinations of collections, machines, and processes that people use to search the collection(s) using the machine(s). Model-based evaluations such as those pioneered at Cranfield and now used in TREC, CLEF, and SIGIR focus on some functions of the machine (in particular, how best to build ranked lists). This paper expands that focus to examine what we have learned about the processes by which those machines will be used to perform Cross-Language Information Retrieval (CLIR), concluding with a brief description of how that perspective informs the nature of our research in DARPA's new GALE program.

Categories and Subject Descriptors

H.1.2 [**Information Systems**]: User/Machine System – *Human Information Processing*

General Terms: Design, Experimentation

Keywords: Interactive Cross-Language Information Retrieval, User-Assisted Query Translation, Experiments, Search Behaviors

1. INTRODUCTION

While it is true that Information Retrieval (IR) technology can be used as a component of some larger system (e.g., clustering, text classification, or question answering), in the normal use of the term "IR" we focus on cases in which the need for a search arises from some human need and the utility of the results will ultimately be judged by the person whose need must be satisfied. In this sense, the dichotomy between "batch" and "interactive" IR is a false one: all IR is ultimately interactive IR. What differs is not what we seek to achieve, but rather what we choose to evaluate. In the Cranfield tradition, we see to determine how well machines can identify documents in a collection that a searcher might wish to see. This abstract formulation covers a broad range of research questions, including query-based topic-oriented ranked retrieval in TREC/CLEF/NTCIR, recent work on query-based sentiment-oriented ranked retrieval in TREC and NTCIR, example-based event-oriented exact-match retrieval in TDT, and evolving simulations of explicit relevance feedback as a basis for exact-match retrieval in the TREC adaptive filtering task and the final year of the TDT topic tracking task. In the Cranfield tradition, we take the information need as fixed (or evolving in some easily modeled way) and we vary the design of the machine.

The research tradition known as "relevance studies" adopts the opposite perspective: the available automated capabilities are (implicitly) taken as fixed, and the research focuses on understanding what factors would cause a user to value the content of a document. The term "relevance" in those studies is used in a broader sense than is typical at SIGIR—closer to what we would normally call "utility." The research methods used to explore this broader notion of relevance are also different from the normal discourse at SIGIR, drawing heavily on cognitive psychology and often relying more on qualitative than quantitative methods of inquiry. The two research traditions intersect in at least one important way: relevance studies quite consistently indicate that the topical relevance that we focus on in the Cranfield tradition is often a dominant factor in the choices made by users (some others are recency, authority, availability, and comprehensibility).

A third related research tradition focuses on the process of asking questions. An example of this that will be familiar to many SIGIR participants is Belkin's "Anomalous States of Knowledge," which observes that we design our machines to answer questions that are well formed, but that those machines are often used by searchers who bring an incomplete understanding of what they are really looking for [1]. Over the years, we have found Taylor's four types of questions (what you really want to know, what you think you want to know, what you can articulate you want to know, and what you can formulate in your machine's query language) to be a useful framework for thinking about this [2]. Two important lines of research emerge from this perspective: the "reference interview" process, and query (re-) formulation strategies. Both have been extensively studied in the context of training information professionals (e.g., reference librarians).

The closest thing that we have to a unifying theory for these three disparate lines of research is sense-making, for which Dervin's iterative situation-gap-bridge is perhaps the best known model [3]. Somewhat oversimplifying in order to draw the connections

SIGIR'06 Workshop, August 11, 2006, Seattle, Washington, USA.
Copyright 2006 ACM 1-58113-000-0/00/0004…$5.00.

clearly, some query formulation process yields a query that, when presented to a machine results in a situation (the search results) that either meets the need or results in perception of a gap. Some or all of Taylor's four types of questions are then adjusted in an effort to bridge that gap, which results in a new situation, from which the process iterates again.

What is missing from all of this is the co-evolution by which changes to one piece of the puzzle result in a cascade of evolutionary improvements. To see how that co-evolution works, we need to step back and re-conceptualize what we mean by an "IR system." In this paper, when we speak of the "system" we mean the collection being searched, the machine that is being used to search it, and the process by which the user uses the machine. When we intend to focus on an automated capability that can be employed by the user, we refer to the machine, not the system.

A simple example should serve to drive home the importance of taking this broader perspective. Google displaced AltaVista as the preeminent Web search engine because Google included several capabilities that AltaVista did not offer at the time. The most often cited advance was the PageRank algorithm, which yielded better ranked lists. But two other features of Google were also important: Google indexed more documents (by indexing anchor text, so that even documents that had not yet been crawled could be found), and Google performed conjunctive (all-terms) queries rather than the disjunctive (any-term) queries that AltaVista and other search engines used at the time. All IR techniques can be thought of as pre-filtering the result set; Google pre-filters with an implicit Boolean *and* across the query terms, earlier search engines had (by returning documents with any non-empty match) implicitly applied a Boolean *or* across the query terms before computing the ranking. The use of conjunctive queries by Google is particularly interesting because it adversely affects many IR measures (e.g., mean average precision and R-precision). Indeed, virtually all Cross-Language IR (CLIR) research today is still based on disjunctive queries for exactly this reason. Google's decision to adopt conjunctive queries was motivated not by getting better results from the machine, but rather by getting better results from the system—a system that includes both the machine and the process by which that machine is used. Evaluating results at this broader scale can be difficult, but in this case the outcome is fairly clear—conjunctive queries are now widely used across the search engine industry. Why? Because they make the search process more effective by enhancing both the searcher's understanding of what the machine has done and their ability to control what it will do next.

Now think for a moment about how users will understand and control the CLIR machines that we have been building for the past decade. First, they need to select some search terms in one language that will be matched with document terms that are written in another. How will they know which search terms our machines can handle? When they get back results, what process will they use to improve their search? Will they add terms that they find in (translated) documents to their query? If they do, will they be re-translated consistently? Will the translations be sufficient to allow them to learn about cultural factors that may change the way they need to formulate their queries, or the way in which they should interpret the results? One way to think of query translation is as cross-language expansion of the query with terms in the document language that have similar meanings. Automatic query expansion is something users may have little experience with (indeed, some Web search engines now even limit the use of stemming). Will users understand enough about how the result set was formed to make an informed guess about how best to modify their query?

The CLEF interactive track (iCLEF) aimed to explore some aspects of interactive CLIR with user studies designed around hypothesis testing (e.g., [4]). For example, we learned that users could often determine topic of retrieved documents (in 2000), that they could often formulate effective queries (in 2002 and 2003) and that they could often find answers to factual questions (in 2004). However, iCLEF studies also pointed up two troubling factors: (1) narrowly focused hypothesis testing was not well suited to answering broader questions about how such collection-machine-process systems should be designed, and (2) very few research teams participated. Indeed, participation in iCLEF never rose to even 10% of overall CLEF participation, and it actually declined over time. Why? Because studying user behavior is expensive, and CLIR researchers make rational decisions about perceived costs and perceived benefits.

Many factors conspire to make the design and execution of interactive studies challenging [5]. Some issues are practical, such as the need for a relatively robust and responsive user interface and the need to recruit a substantial number of study participants with backgrounds that are representative of the intended user population. Indeed, early in the development process there may not be any users with the requisite experience because there are not yet any deployed machines with which they could have gained that experience! Other issues impose fundamental limits on the ability to measure the effectiveness of our designs, most notably the variance introduced by individual differences between study participants that are difficult to control for (e.g., prior knowledge or reading ability) and by differing understandings of the assigned task, and the difficulty of standardizing experiment conditions across multiple participating sites. Fatigue and learning effects are also significant factors that must be accommodated in a study design, particularly when attempting to study iterative search processes that may extend over multiple sessions. Finally, our inability to anticipate some of these issues further compounds the challenge. Few of us get it right the first time we run a batch experiment with a new collection. In that case, the cost of a replication after fixing the problem is fairly modest. Not so, however, for user studies. The upshot of all this is that controlled quantitative user studies are an expensive way of learning about how people will use the machines that we build.

So we are faced with a conundrum: if we are to design useful machines, we must understand the process(es) by which those machines will be used. But the user studies we have been doing have at best a limited potential to help us gain that understanding. Clearly, we need to explore a broader range of approaches. In this paper, we propose one such strategy. We start in the next two sections by revisiting our 2002 and 2003 iCLEF experiments, focusing on we learned about the process rather than the results. We then use this to motivate our planned study for a new project in which we are collaborating with a multidisciplinary team to iteratively evolve the design of machines to support this task.

2. THE MIRACLE USER STUDIES

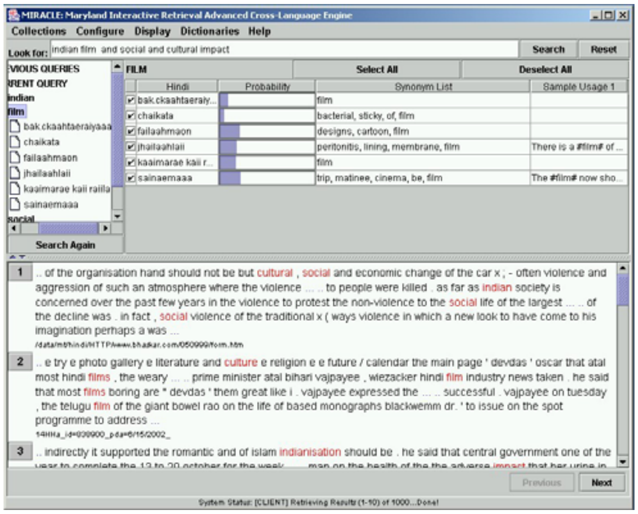
MIRACLE (the Maryland Information Retrieval Advanced Cross-Language Engine; see Figure 1) is an interactive query translation based CLIR application designed to support rapid prototype iteration, and to explore interaction design for interactive CLIR [6]. In MIRACLE, users have four interactive points to support refinement of their mental models of their needs, the machine's capabilities, and the collection. Three of these, *query formulation*, *document selection from search results*, and *document examination*, are familiar from monolingual applications such as Web search engines. The fourth, *query translation*, is unique to CLIR. Our approach to query translation in MIRACLE is to take advantage of the presence of the users, inviting them to participate to the process of constructing a document-language query based on the source-language query terms that have been entered. We call this approach "user-assisted query translation" [7]. It is designed to foster transparency and control, facilitating the searcher's development of mental models of the machine's operation. Selecting correct translations could improve results, although omitting a useful translation could equally well have an adverse effect. Our principal motivation for including this capability was to support iterative query refinement: if users make bad choices, they can see the effect and learn to better control the machine.

Three types of cues are provided to help monolingual users determine which translations should be selected: (1) the translation itself (transliterated if necessary), the meaning of which might be recognized by the user if it is a loan word or a proper name; (2) a list of possible synonyms (found using a "back translation" technique that relies on the presence of near-synonyms among the translations of a term [8]); and (3) examples of usage (found in a separate set of translated or topically related texts). These cues were all generated automatically; more information on how they are created can be found in [5].

The design of MIRACLE was shaped by two key design guidelines: (1) expose our interaction design to the user in a straightforward and easily understood manner, and (2) provide immediate feedback in response to control actions. These both contribute to the overarching design goal of MIRACLE: to support the progressive refinement of mental models that can contribute to improved search effectiveness.

The data presented in this paper are from three experiments (two in April 2002, and one in April 2003) that we conducted using variants of a single study design. All three experiments used a within-subjects design, where each subject (i.e., searcher) performs repeated trials (several searches, each for a different topic). The order of those trials was varied systematically in order to block (average out) the effects of presentation order on learning and fatigue, the effects of individual differences in users, and the effect of differences in topic difficulty [9]. Each study was designed to compare two conditions, the user-assisted condition (using the full capabilities of user-assisted query translation in MIRACLE) and the automatic condition (the same interface, but with the query translation and translated query display areas permanently hidden).

Because the query language of MIRACLE is English, we chose document languages other than English. For studies 1 and 2, we elected to work with the CLEF German document collection, which contained 71,677 news stories from the Swiss News Agency (SDA) and 13,979 news stories from Der-Spiegel. For study 3, we used the CLEF Spanish document collection, which contained 215,738 news stores from the EFE News Agency. In each case, we automatically translated the documents into English using Systran Professional 3.0 to support construction of summaries (for display in a ranked list) and for display of full document translations (when selected for viewing by the user).



Look for: indian film and social and cultural impact

	Hindi	Probability	Synonym List	Sample Usage 1
☑	bak.ckaahtaeraly...		film	
☑	chaikata		bacterial, sticky, of, film	
☑	failaahmaon		designs, cartoon, film	
☑	jhailaahlaii		peritonitis, lining, membrane, film	There is a #film# of ...
☑	kaaimarae kaii r...		film	
☑	sainaemaaa		trip, matinee, cinema, be, film	The #film# now sho...

1 .. of the organisation hand should not be but cultural , social and economic change of the car x ; - often violence and aggression of such an atmosphere where the violence ... .. to people were killed . as far as indian society is concerned over the past few years in the violence to protest the non-violence to the social life of the largest ... .. of the decline was . in fact , social violence of the traditional x (ways violence in which a new look to have come to his imagination perhaps a was ...
/data/mt/hindi/HTTP/www.bhadkar.com/050999/form.htm

2 .. e try e photo gallery e literature and culture e religion e e future / calendar the main page ' devdas ' oscar that atal most hindi films , the weary ... .. prime minister atal bihari vajpayee , wiezacker hindi film industry news taken . he said that most films boring are " devdas ' them great like i . vajpayee expressed the ... .. successful . vajpayee on tuesday , the telugu film of the giant bowel rao on the life of based monographs blackwemm dr. ' to issue on the spot programme to address ...
14HHa_id=030900_pda=6/15/2002_

3 .. indirectly it supported the romantic and of islam indianisation should be . he said that central government one of the year to complete the 13 to 20 october for the week ... man on the health of the the adverse impact that her urine in

System Status: [CLIENT] Retrieving Results (1-10) of 1000...Done!

Figure 1: MIRACLE interface with the user-assisted query translation panel displayed.

A total of 12 CLEF topics were used in three studies. The topics were provided to the participants as a written topic statement that consisted of a title field, (typically in a keyword-oriented telegraphic style), a description field (which could be thought of as representing what a user might initially say to someone who was helping them with their search), and a narrative field (providing additional information about how to distinguish between relevant and irrelevant documents). Our participants iteratively constructed queries based on their understanding of the topics and the observed behavior of MIRACLE.

Participants searched individually, which made it possible to collect the observational data reported below. Sessions took about 2.5 hours per participant. There was a fixed time for each topic (10 minutes or 20 minutes, depending on the study), plus training, breaks, questionnaires, and interviews.

3. THE CLIR SEARCH PROCESS

Our goal in this paper is to draw on our experience in these three user studies to gain insights into the way our participants actually used an interactive CLIR application. The data used for this analysis were obtained from several sources, including queries and related activities (e.g., translation selection) for each iteration, recorded screen captures for each search session, and an exit interview conducted at the end of each session. While our study context, designed as it was principally to support quantitative hypothesis testing, resulted in studying searchers who were performing stimulated rather than self-directed situated tasks, the qualitative analysis we report on here otherwise resembles what would be done in a case study setting.

3.1 Participants

We had a total of 20 participants in our experiments; no person participated in more than one study. The population was relatively homogeneous:

- *Native English speakers, generally with little or no proficiency in the document language*. All 20 participants were native speakers of English. Nineteen reported either no reading skills or poor reading skills in the document language (German or Spanish); one participant (in study 1) reported good reading skills in the document language (German).
- *Inexperienced with machine translation*. Eighteen of 20 participants reported never having used any machine translation software or Web translation services. The remaining two reported "some experience" with machine translation software or services.
- *Experienced searchers*. Eleven of the 20 participants had received formal education in library science (generally as current library science students). The participants reported an average of about 7 years of on-line searching experience, with a minimum of 3 years and maximum of 10 years. Most participants reported extensive experience with Web search services, and all reported at least some experience searching computerized library catalogs (ranging from "some" to "a great deal"). Almost all (19 of 20) reported that they search at least once a day.
- *Highly educated*. Sixteen of the 20 were either currently enrolled in a program leading to a Masters degree or had already earned at least a Masters degree. The remaining four had either completed or nearly completed a Bachelors degree.
- *Mature*. The average age over all participants was 32, with the youngest being 21 and the oldest being 45.
- *More often female*. There were 13 female participants and 7 male participants.
- *Not previous study participants*. None of the participants had previously participated in a TREC or iCLEF study.

3.2 Search strategies

A "search strategy" refers to a plan that a user constructs to guide their search process [10]. Marchionini identified several common search strategies, including formal techniques in which librarians are trained (e.g., pearl growing, successive fractions (onion peeling), and building blocks) and emergent strategies (e.g., "interactive easy search") that end users of search engines seem to naturally develop without formal training [11]. A hallmark of Marchionini's "interactive easy search" strategy is reliance on immediate access to full text, from which both new concepts and new vocabulary can be iteratively acquired.

All participants were observed to use some variant of this "interactive easy search" process, either alone or in combination with other strategies, in most of their searches. This may result from the fact that the participants did not know much about MIRACLE's design, the collection, or (in many cases) the topic before beginning their search. The prior knowledge of the topic is to some extent an artifact of our study design (since we, rather than they, chose the topics), so this result should be interpreted with caution.

Variants of the "building block" strategy, in which separate sub-queries are constructed for each facet of an information need, were also observed in our studies. One topic required searchers to find documents on two facets of the marriage between Michael Jackson and Lisa Marie Presley: their wedding, and their subsequent separation. In this case, most participants (3 library science students and 3 others, out of a total of 8 participants in that study) employed a building blocks variant. The building blocks strategy taught to librarians results in construction of a single query in conjunctive normal form *(and* across facets, with nested *or* across facet-specific vocabulary). MIRACLE does not support Boolean queries, so participants first searched for documents on one facet of the topic, then for the other. We also identified four other cases in which a variant of the building blocks strategy was employed for a topic where the potential benefit of facet-specific searching was less immediately obvious. In every one of those cases, the participant was a library science student. From this we conclude that professional searchers may employ CLIR applications in ways that are different from what experiments with other types of searchers would lead us to expect.

3.3 Sources of query terms

The majority of terms in the initial queries issued by our participants were present in the topic statements that we provided. Participants were also observed to initially select terms from their own prior knowledge about a topic. For example, one participant who happened to be an expert on computer security included the term "intrusion detection," which was not in the topic statement. In another example, one participant used "CGI" and another used "pixar" in their initial query for a topic about computer animation in films because they had adequate background knowledge on the topic. The third source of terms for the initial query was linguistic knowledge of synonymy, abbreviations and morphological variants. For example, "American America" appeared in an initial query when "U.S." had been in the topic statement, "anti-racism anti-prejudice" appeared in a query for a topic in which the topic statement contained "against racism," and one initial query was "computer animation animations film films movie movies."

There was no noticeable difference between the selection of initial query terms between the automatic and user-assisted conditions. There were, however, clear differences in subsequent search behavior between the two conditions. We focus on the user-assisted condition in detail below (in section 3.4), so in this section we focus on the automatic condition. Participants behaved conventionally according to Marchionini's "interactive easy search" process, adopting terms from relevant documents, adding or removing terms from their query, using synonyms or hyponyms (more specific terms), etc. This is not surprising, since our automatic condition was designed to replicate as closely as possible the functions provided by a typical Web search engine.

Interestingly, there was one case in which the participant chose to add a document-language term to the query, apparently based on guessing from context that it might be a useful query term. In that case, a search for information about computer animation, the participant added "king Leon" to the query, probably because Systran had failed to translate the last word in ``El rey Leon'' ("Lion King") when it appeared in a document. This resulted in finding several additional relevant documents because Leon was (fortunately) an untranslatable term that MIRACLE passed through unchanged. That incident suggests that intentionally incorporating facilities for document-language feedback might be useful in some cases, and that the handling of untranslatable terms should receive specific consideration when designing interactive CLIR applications.

3.4 User-assisted query translation process

Our analysis identified several ways in which searchers sought to exploit the new capabilities that our user-assisted query translation feature offered. While much of what we saw overall was similar to what we observed in the automatic condition, some of our participants proved to be delightfully inventive in the limited time that they had to work with MIRACLE. We observed four new strategies (listed here in decreasing order of prevalence):

- *Translation selection and deselection*. In two of our three studies, every participant did actually try deselecting at least one unwanted translation at some point in their session based on the cues that MIRACLE provided (all translations were selected by default). On average in these two studies, 23% of the search iterations involved either explicit translation deselection or reselection. In some cases, participants returned repeatedly to change their choices from among the available translations. Two patterns of use were observed, sometimes separately, but often combined:
    - **Query-Translate-Search:** The searcher issued a query, performed translation selection/deselection in the translation panel, then clicked the search button to request documents.
    - **Search-Translate-Search:** The searcher obtained a set of returned documents after clicking the search button, they examined translated document snippets and/or translated documents, they then went to the translation panel to select/deselect translations, and then they clicked the search button to request another set of results.

  It is hard to know how much of this observed behavior resulted from exploration to learn how the new capability worked, what part resulted from using it because of its perceived utility, and what part resulted simply from playing around with something new. A longitudinal study would be needed to determine whether searchers continued to use this capability once the novelty wore off and they had more experience with it.

- *Assessing the utility of a query term*. We also observed a **Query-Translate-Query** pattern in which the searcher issued a query, examined the available translations in the translation selection panel, and then decided to change part or all of their initial query before performing a search. For example, during a computer animation search, one participant first entered the query “movie film computer animation CGI.” They then removed several unwanted translations, but before clicking the “search” button changed the query by replacing “animation” with “animated.” They then examined the known translations for “animated,” and changed the query term back to “animation.” Clearly, that searcher was using the information gained in the translation selection panel as the indicator to the potential utility of query terms. We observed similar behavior from several other participants; about 18% of all query iterations involved this kind of behavior. From this we conclude that searchers sometimes gain a greater degree of insight into the behavior of the machine that they are using when user-assisted query translation is available.

- *Vocabulary selection based on translations, back translations, or examples of usage*. In several cases, we observed that the terms added into search queries were not from returned documents, but from the translation selection panel. In the most blatant example, after posing several queries that contained variants of ``European Union," one participant simply selected one of the displayed Spanish translations for each word (i.e. “europeo” for European and “sindicato” for union) and typed them directly into the query. MIRACLE treated both as untranslatable words, and the participant was able to find two additional relevant documents based on that query. Interestingly, that participant used the same trick several times when they needed European Union in queries for subsequent topics.

- *Translation-based spelling verification*. MIRACLE highlights query terms that have no known translations by showing the term in red in the translation selection panel; such terms are retained unchanged. This feature was originally included so that participants could use their domain or linguistic knowledge to replace unknown terms with some synonym for which translations were known. We observed, however, that some participants found that this feature was also helpful for detecting spelling errors (since misspelled words will typically have no known translation). For example, one participant twice noticed misspellings in their queries, quickly correcting “policy” to “policy,” and “preley” to “presley.” It is well known in other contexts that users appropriate new technology and use it in unexpected ways. Only by observing people using our machines can we begin to appreciate the implications of this for our designs.

3.5 Factors affecting translation selection

User-assisted query translation was used more often when more time was available. On average across the three studies, 30% of all search iterations were preceded by one or more translation deselection or reselection actions. In the first two studies, with 20-minute search sessions, the average was 40%, whereas in the third study, with 10-minute search sessions, the average was only 18%. Moreover, in the 20-minute sessions we observed that participants performed translation deselection or reselection more often in the second half of their session than that in the first half (55 times vs 34 times). From these observations we conclude that our participants found some utility in the fully automatic feature, generally turning to the user-assisted query translation capability if time remained after exhausting what could be found using simpler techniques. This suggests that our present design, in which users can easily hide the translation selection window to gain more screen space for the automatic search results (and can easily restore it later), is well suited to the way in which MIRACLE was actually employed.

Topic difficulty (indicated by a relative paucity of relevant documents in the collection) also seemed to affect the use of user-assisted query translation. Because of a quirk in our study design, the collection being searched in the study 2 was a strict subset of the collection searched in study 1.[1] Although the collection in study 2 was about 2/3 the size (about 14,000 vs. about 21,000 documents), the number of relevant documents for each topic was far smaller in study 2 (averaging 11 with a range of 0-21 in study 2, compared with an average of 47 with a range if 22-68 in study 1). In study 1, 47% of the search iterations were preceded by one or more translation deselection or reselection operations. For study 2, with far fewer relevant documents, this dropped to an average of 34%. Interestingly, the drop can be entirely explained by less use of the Query-Translate-Search and Search-Translate-Search patterns (from 30% in study 1 to 15% in study 2), whereas use of the Query-Translate-Query pattern actually increased slightly from to 17% to 19%. We interpret this as an additional source of support for our conclusion that searchers actually do find new query terms in translated snippets and translated documents. This has important implications for the degree of integration between the translation techniques used for presentation of results and the implementation of query translation capabilities. In our present implementation of MIRACLE these are completely independent. For an operational application, there is now clear evidence that some form of closer coupling would be warranted.

4. DISCUSSIONS

Several profound conclusions can be obtained from the studies reported above. First, studying the process(es) by which CLIR machines are used is as important as examining the effectiveness of those machines in producing desired results. All interesting and important insights presented in this paper are from examining the actual behaviors of our users, which would not be revealed if we only look at the performance results. The second conclusion is that participants' behaviors are changing even during the short time span of the experiments. Participants learned from the interactions, adapted to the capabilities of the machines, and developed their own tactics to explore the situation. Therefore, the design of CLIR machines should aim at helping people to quickly develop their tactics, and the evaluation design for CLIR machines should take participants' adaptation into consideration. The third conclusion is that there is no typical users because users are different. Users' experience, skills, and background all cannot only influence their performance, but also affect their behaviors with the CLIR machines. Combine the second and the third conclusions, there is also a call for longitudinal study of users working with CLIR machines. Only in a long time period, will users' behavior and their tactics/strategies can be evolved and be captured in the study.

Our CLEF user studies were, however, limited in many ways by the constraints of the quantitative hypothesis testing that had been our primary focus at the time that the studies were designed. There is much more that we need to learn, and additional studies will be needed before we will be in a position to characterize the degree to which these results would generalize to other user groups, non-topical search tasks, or settings in which documents are available in more than one language. Addressing that challenge calls for a shift in our thinking, from studies in which observational methods were intended to enrich hypothesis testing to a new group of studies designed to make the most of what we can learn from observation of situated users. That is the key idea shaping the design of our planed user studies for DARPA's new GALE program.

5. THE GALE USER STUDIES

Looking back over the past decade, three driving forces behind research on CLIR are apparent. The first to emerge was the World-Wide Web, the very name of which evokes a vast multilingual commons in need of this technology. Although isolated projects were undertaken at Cornell in the early 1970's and at Bellcore around 1990, the spontaneous emergence of a global community of researchers coincided with the emergence of the Web between roughly 1992-1994 (when several projects began) and 1996 (when we first met at SIGIR in Zurich). A second driving force was the emergence of a global set of evaluation venues, initially at TREC (and to a lesser extent, at TDT), and ultimately at CLEF and NTCIR. This focus on evaluation was complemented by a third key enabler, nearly simultaneous decisions by several funding agencies to make substantial sustained investments in basic research on CLIR. Each of these forces drives the community in somewhat different directions. In this section, we follow the money, describing where work funded by what has to date been one of the world's largest sponsors of CLIR research, the Defense Advanced Research Project Agency (DARPA), is headed.

DARPA's investment in CLIR began in 1997 with an initial exploratory effort as part of a broader Information Management (IM) program. Starting in 2000, DARPA vastly expanded its investment in language technology with the Translingual Information Detection, Extraction, and Summarization (TIDES) program [12]. As the name suggests, TIDES was structured to advance the state of the art in four key component technologies: semantic tagging in multiple languages for enhanced content representation, CLIR, summarization of multilingual content from single and multiple documents, and statistical machine translation for presentation of retrieved information. DARPA's new Global Autonomous Language Exploitation (GALE) program continues this work, with an additional focus on speech transcription. But the most remarkable feature of GALE is a strong focus on creation and evaluation of fully integrated systems.

Figure 2 presents one way of looking at a GALE architecture. The acquisition component shown at the left encapsulates the technology that essentially serves to convert spoken or written language into a computationally manipulatable sequence of terms. The extraction component then enriches that term sequence with annotations to permit richer content modeling. The presentation component operates on the resulting representations, snipping passages, selecting which should be shown, and synthesizing some form of summary from which the user can optionally drill down to see the original (translated) source. This process must be guided by some model of what the user would wish to see. It is not hard to see how a present text-oriented CLIR application would map onto this model: acquisition is simply tokenization (or perhaps tokenization followed by term-by-term translation),

[1] Actually, this happened in the opposite order. Study 2 had been intended as our iCLEF 2002 submission, but after completing the study we discovered that we had failed to index a part of the collection. We therefore fixed the error and reported our replication with the full collection as "study 1" because it was our official submission. We continue to use those names in this paper for consistency with our previous publications.

extraction is omitted, presentation is document boundary detection and document ranking, and the (possibly translated) query is the entirety of the user model (which the user inputs as the "control" action). But GALE offers a far more extensive sandbox with which to explore possibilities: integration of state-of-the-art entity recognition to improve name translation, various ways of augmenting the result set in language-independent ways (e.g., with social network visualizations), and richer interaction models.

Making the best use of this unique opportunity will require that we step back to look broadly at how these capabilities should be used, and to determine which will prove to be the most useful. We therefore plan to conduct an extensive series of formative user studies over the next year to begin to explore the design space. Our colleagues at the IBM T.J. Watson Research Center already have an initial CLIR application running into which we and our partners can integrate components, and by the time we meet in Seattle for this SIGIR workshop we will quite likely have some initial study results.

The GALE user study provides us several unique opportunities. First, GALE identifies specific user groups and employment scenarios, giving our experiment design concrete targets to work on. Second, as shown in Figure 2, CLIR application is only one part of a whole integrated information system. Comparing to a stand-alone CLIR machine, this is probably closer to the real application scenarios of CLIR in current situations. In addition, the judgment on CLIR's performance is not precision and recall in abstract, but whether or not it can deliver reasonable quality cross-language results to let subjects complete their tasks. Third, GALE machines will evolve over a period of several years, providing the possibility of a long term study of the machines and the users who work on them.

Briefly, our user study is formative end-to-end evaluation. We will examine the whole GALE machine rather than individual components. We have identified several employment scenarios, including:

1. One analyst, compiling a descriptive report;
2. One user trained both in information gathering and decision making, compiling a report recommending a course of action;
3. One decision maker collaborating face to face with one or more searchers, each of whom is supported by a GALE machine.

The last scenario is characterized by informal interaction between the decision maker and the searcher(s) that continues in cycles until a decision is made. By assigning decision making and searching responsibilities to different participants, we are in effect externalizing some aspects of the thought processes of the user in the second scenario. We are testing several types of tasks within each employment scenario. The output of all the testing tasks is a short paper. Tasks are based on lessons learned from previous studies and typical intelligence analysis tasks, e.g. tracking a topic over a several week span. The topics used in the study are drawn from current events and selected for their appropriateness to the employment scenario. The data are live text and video feeds in three languages. We recruited library science students, who are professional searchers, as surrogates for analysts, and similarly situated surrogates for trained decision makers (e.g., retired military officers). We employ interviews, log analysis, and observational notes to construct an understanding of what works, what doesn't, and where there are opportunities for improvement.

While this will not be the first such effort, the GALE program offers an unprecedented opportunity to bring together state of the art technology from a broad array of relevant disciplines. If we are to make the most of this remarkable opportunity, we must think about much more than how to construct better ranked lists—we need to think broadly about the design and evaluation of fully integrated multilingual information systems.

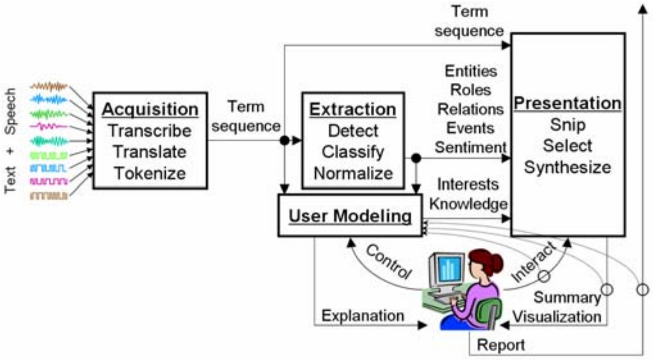


Figure 2. Proposed GALE architecture

6. RELATED WORK

Our CLEF user studies presented in this paper were not unique in the literature. Participants in CLEF interactive track have been examining various issues related to interactions and users in CLIR for several years. For overview of the activities, please read Gonzalo and Oard's iCLEF oview papers since 2001 (e.g.,[4]). Most of the reported research has been concentrated on the effectiveness of certain techniques in improving users' performance in CLIR applications. For example, He et. al. confirms the usefulness of user-assisted query translation [7], López-Ostenero et. al. demonstrates the effectiveness of using phrase translation in CLIR [13], and Petrelli et. al. identify the importance of back translation selection/de-selection in CLIR interaction [14]. However, there are qualitative analyses of the CLIR interactions too. For example, the using of target language terms in queries has been discussed in reports from various sites [15, 16], and the effect of mistranslations/missing translations is presented in CLEF 2005 [16].

7. CONCLUSION

A decade of research on cross-language information retrieval has yielded effective and efficient techniques for ranking documents in one language based on queries and/or examples that are expressed in another, but technology to support other aspects of interaction with the user is not yet very mature. Incremental progress in cross-language ranking techniques will undoubtedly continue, but it is becoming increasingly evident that if we want the techniques that we develop to actually be used we need to take a broader view. When we met to consider the future of the CLIR community in 2002, this point received extensive discussion [17]. In this paper, we have looked back at what we have learned since then, and started to sketch out what now seems to be a good direction to head.

8. ACKNOWLEDGMENTS

The authors are grateful to Jianqiang Wang and Jun Luo for their help with the user studies reported in this paper. This work has been supported in part by DARPA contracts N660010028910 (TIDES) and HR0011-06-2-0001 (GALE).

9. REFERENCES

1. Belkin, N.J., *Anomalous states of knowledge as a basis for information retrieval.* Canadian Journal of Information Science, 1980. **5**: p. 133-143.

2. Taylor, R.S., *Question-negotiation and information seeking in libraries.* College & Research Libraries, 1968. **29**: p. 178-94.

3. Dervin, B., *Sense-Making theory and practice: An overview of user interests in knowledge seeking and use.* Journal of Knowledge Management, 1998. **2**(2): p. 36-46.

4. Gonzalo, J. and D.W. Oard. *The CLEF 2002 Interactive Track* in *Proceeding of Cross-Language Evaluation Forum 2002*. 2002.

5. Dumais, S.T. and N.J. Belkin, *The TREC Interactive Tracks: Putting the User into Search*, in *TREC: Experiment and Evaluation in Information Retrieval*, E.M. Voorhees and D.K. Harman, Editors. 2005, MIT Press. p. 123-152.

6. He, D., et al., *Making MIRACLEs: Interactive Translingual Search for Cebuano and Hindi.* ACM Transactions on Asian Lnaguage Information Processing, 2003. **2**(3): p. 219-244.

7. He, D., et al. *Comparing user-assisted and Automatic Query Translation*. in *Proceedings of CLEF'02*. 2002.

8. Capstick, J., et al., *A system for supporting cross-lingual information retrieval.* Information Processing and Management, 2000: p. 275-289.

9. Maxwell, S.E., H.D. Dalaney, and J.W. Dimmick, *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. 2003: Lawrence Brhaum Assoc.

10. Bates, M.J., *Information Search Tactics.* Journal of the American Society for Information Science, 1979. **30**(4): p. 205-214.

11. Marchionini, G., *Information Seeking in Electronic Environments*. 1995: Cambridge University Press.

12. Oard, D.W. and A.L. Powell, *Team TIDES Newsletter*. 2002-2005.

13. López-Ostenero, F., et al. *Interactive Cross-Language Searching: phrases are better than terms for query formulation and refinement*. in *Proceedings of Cross-Language Evaluation Forum*. 2002.

14. Petrelli, D., et al., *Which user interaction for cross-language information retrieval? Design issues and reflections.* Journal for American Society of Information Science and Technoloqy, 2006. **57**(5): p. 709-922.

15. Dorr, B.J., et al. *iCLEF 2003 at Maryland: Translation Selection and Document Selection*. in *Proceedings of CLEF'03*. 2003.

16. Petrelli, D. and P. Clough. *Concept Hierarchy across Languages in Text-Based Image Retrieval: A User Evaluation*. in *Proceedings of Cross-Language Evaluation Forum*. 2005.

17. Oard, D.W. *When You Come to a Fork in the Road, Take It!* in *Proceedings of SIGIR2002 workshop "Cross-Language Information Retrieval: A Research Roadmap"*. 2002.