Available online at www.sciencedirect.com

ScienceDirect

Information Processing and Management 44 (2008) 181–211

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

User-assisted query translation for interactive cross-language information retrieval

Douglas W. Oard [a,*], Daqing He [b], Jianqiang Wang [c]

[a] College of Information Studies and Institute for Advanced Computer Studies,
University of Maryland, College Park, MD 20742, United States

[b] School of Information Sciences, University of Pittsburgh, 135 North Bellefield Av, Pittsburgh, PA 15260, United States

[c] Department of Library and Information Studies, University at Buffalo, State University of New York, 528 Baldy Hall,
Buffalo, NY 14260, United States

Received 12 September 2006; received in revised form 21 December 2006; accepted 22 December 2006
Available online 7 March 2007

Abstract

Interactive Cross-Language Information Retrieval (CLIR), a process in which searcher and system collaborate to find documents that satisfy an information need regardless of the language in which those documents are written, calls for designs in which synergies between searcher and system can be leveraged so that the strengths of one can cover weaknesses of the other. This paper describes an approach that employs user-assisted query translation to help searchers better understand the system's operation. Supporting interaction and interface designs are introduced, and results from three user studies are presented. The results indicate that experienced searchers presented with this new system evolve new search strategies that make effective use of the new capabilities, that they achieve retrieval effectiveness comparable to results obtained using fully automatic techniques, and that reported satisfaction with support for cross-language searching increased. The paper concludes with a description of a freely available interactive CLIR system that incorporates lessons learned from this research.

Keywords: Cross-language information retrieval; User studies; Machine translation

1. Introduction

Cross-Language Information Retrieval (CLIR) systems seek to identify topically relevant documents that are written in one language (e.g., French) based on queries that are expressed in another (e.g., English). Early CLIR systems were designed to identify an unranked set of documents based on Boolean queries and a multilingual thesaurus (Oard & Diekema, 1998). Over the past 15 years, however, ranked retrieval based on ‘‘natural language’’ queries has become the dominant paradigm for CLIR research. Retrieval results

[*] Corresponding author. Tel.: +1 301 405 7590.
E-mail address: oard@umd.edu (D.W. Oard).

0306-4573/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2006.12.009

can be used to help searchers: (1) recognize relevant documents to be used outside of the retrieval system, (2) gain insight into the way the system operates so that they can express their information needs in ways that will result in effective searching, and (3) better understand the true nature of their information needs. In an earlier study, we found that present machine translation systems can provide a useful degree of support for the cross-language document recognition task (Oard, Gonzalo, Sanderson, Lopez-Ostenero, & Wang, 2004). The focus of this paper is therefore on the second challenge; we seek to design systems that facilitate query formulation and reformulation for CLIR systems that employ ranked retrieval.

Searching for information is, ultimately, a human activity. Humans and machines can bring complementary strengths to an interactive search process; properly coupling these capabilities can result in a synergy that exceeds the ability of either human or machine alone. Fig. 1 illustrates the close coupling between system design and the process by which the system will be used. Search strategies learned through experience or formal training guide the user's interaction with the search system. For example, librarians learn to use facet analysis to formulate Boolean queries in conjunctive normal form (Marchionini, 1995), resulting in greater success than is typically observed when untrained searchers employ Boolean search systems. Web searchers with no specific training in search strategies have nevertheless also been observed to employ systematic techniques for exploring alternative query formulations (Spink & Jansen, 2004). This poses a co-design problem: search systems must support the processes that searchers actually employ, but new capabilities (e.g., CLIR) can inspire the development of new processes. Iterative prototype refinement can be a useful approach to requirement elicitation in such cases. In this paper we present formative user study results for three prototype iterations and describe the resulting system design.

Because our principal focus here is on supporting query formulation, we sought to provide the searcher with insight into how their query terms are being matched with terms found in the documents. This led us to select an architecture based on translating the query terms into the document language. We call the resulting process *user-assisted query translation*. In this paper, we seek answers to the following broad questions:

- How should support for user-assisted query translation be designed? We explored this question by integrating three techniques for exploiting available bilingual resources (e.g., dictionaries and corpora), ultimately coupling them with an architecture that supports progressive refinement.
- How will searchers employ user-assisted query translation capabilities? We sought evidence for this by using mixed-method user studies in which quantitative comparisons were augmented with rich collection of observational data.
- What is the effect of introducing user-assisted query translation on search outcomes (e.g., retrieval effectiveness and user satisfaction)? We conducted three formative studies with a total of 20 users that begin to characterize the effects of specific system design decisions on representative users performing realistic tasks.

In Section 2, we describe the system design that supported our user studies, first reviewing the extensive work on fully automated techniques for CLIR and the research to date on interactive CLIR, and then describing the design of a flexible interactive CLIR system that was intended to support iterative prototype refinement. Section 3 then reports what we learned from three user studies with variants of that system. Section 4 draws on those results to inform the design of a new interactive CLIR system that we are making freely available to support further experimentation. The paper concludes with a discussion of future directions for research on interactive CLIR.

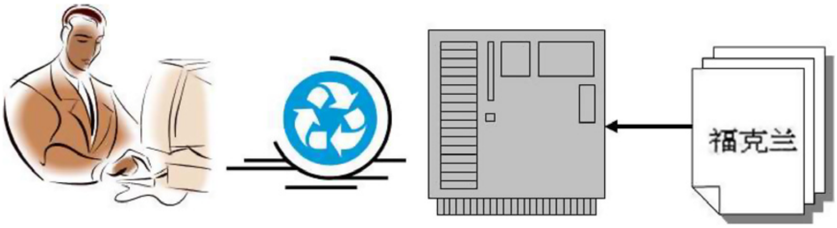


Fig. 1. Illustrating the mutual dependence between search system design and the process by which that system will be used.

2. Interaction design for CLIR

Modern research on CLIR dates from 1990, when Landauer and Littman first explored the potential for large-scale ranked retrieval of documents in one language using queries expressed in another (Landauer & Littman, 1990). Over the course of the next decade, this problem was essentially solved: experimental studies now routinely report ranked lists built using CLIR systems that are nearly as good (by typical ranked retrieval measures) as ranked lists built using same-language queries(e.g., Peters et al., 2004). Of course, this raises two additional questions: (1) where do the queries come from?, and (2) what will the searcher do with the documents that they find? Several answers have been proposed to the second of those questions, including (Oard, 2002):

- It might suffice to know that a document exists (e.g., when seeking to learn who is working in a field that is new to the searcher).
- Documents that appear to be relevant can be submitted to professional translation services. Experiments with human subjects indicate that existing translation technology is often adequate to support the document selection task.
- A text-based search might serve as a basis for finding related content that does not require specific language skills (e.g., images or instrumental music).

The question of where the query comes from might seem straightforward—of course, it comes from the searcher. But saying that begs the question of how the searcher learned to formulate the right query. Searchers often find over the course of a search session that their understanding of what they are actually looking for is incomplete. Moreover, they may also need to learn to effectively express those information needs. Strategies based on iterative refinement are commonly used in such cases (Marchionini, 1995). The success of iterative refinement depends on two types of knowledge: an understanding of why the machine produced the results that were obtained, and an understanding of the ways in which the outcome could be altered. Searchers can therefore be viewed as seeking to refine three mental models: (1) their understanding of their own information need, (2) appropriate query terms that might be present in the documents that are sought, and (3) ways of combining these terms to best express the need (i.e., the "query language"). In this paper, we focus mainly on the query term selection process because it is that process that distinguishes cross-language search from its monolingual counterpart.

Searchers can leverage feedback to support refinement of their mental models. Fig. 2 illustrates four interaction opportunities. Three of these, *query formulation*, *document selection* from a ranked list, and *document examination*, are familiar from monolingual applications such as Web search engines. The fourth, *query translation*, is unique to interactive CLIR. Our approach to query translation is to take advantage of the presence of the searcher, inviting them to participate the process of constructing a document-language query based on the source-language query terms that they have entered.

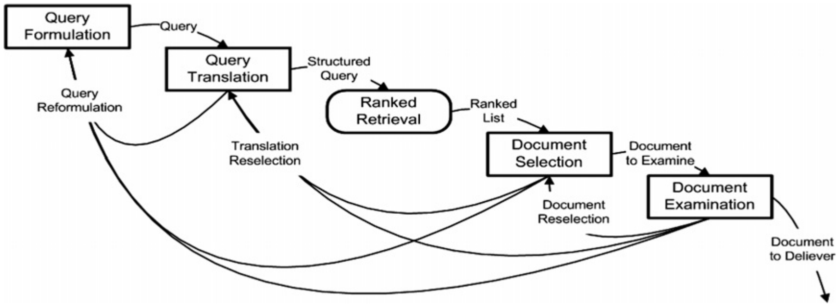


Fig. 2. The four possible interaction points in interactive CLIR: query formulation, query translation, document selection, and document examination. Ranked retrieval is a fully automatic process.

To achieve synergy between searcher and system, we must understand the system's capabilities. Two capabilities are important in this regard: (1) the system's ability to find documents that the user might wish to see, and (2) the system's ability to explain to the searcher how their choices will affect the retrieval results. We therefore begin by briefly surveying what is known about automatic and interactive CLIR, and then return at the end of this section to the development of a comprehensive model for supporting iterative query refinement.

2.1. CLIR techniques

System architectures for CLIR can generally be classified as query translation, document translation, or interlingual (Oard & Diekema, 1998). We have adopted a dictionary-based query translation architecture for the work reported in this paper because the query translation process can be crafted in ways that place the locus of control with the user. CLIR systems that use dictionary-based query translation face three key challenges: (1) selecting query terms for translation in ways that are compatible with the available bilingual dictionaries and that accurately reflect the searcher's intended meaning, (2) using the known translations found in the bilingual dictionaries in ways that optimize retrieval effectiveness, and (3) accommodating cases in which no translation is known for a query term. Each provides interaction opportunities.

Bilingual dictionaries are normally organized to provide translations for root forms of individual words (e.g., "*laugh*" rather than *laughing*"), so term selection for CLIR must accommodate cases in which only a portion of a query term can be found in the dictionary. One commonly used strategy when no translation is known for a query word is therefore to back off to the root form of that word and to use that root as a basis for translation. Stemming the query word (and all single words on the source-language side of the dictionary) offers a robust and easily implemented alternative to full morphological analysis (Resnik, Oard, & Levow, 2001). This can, of course, result in generation of incorrect morphological variants, but document-language stemming limits any adverse effect from that factor on retrieval effectiveness.

Much of the research on CLIR has focused on the cases in which more than one translation is known for a query term. Ambiguity is an unavoidable consequence of using natural language, but CLIR applications must accommodate ambiguity in both the query language and the document language. In monolingual applications, interactive search systems can accommodate word sense ambiguity by allowing the user to group words (typically by using quotation marks) into multi-word terms that must appear together and in order. Some fully automatic CLIR systems achieve a similar effect by reversing the translation dictionary so that multi-word terms appear on the source-language side of the dictionary and then using greedy maximum-length sequence matching to identify multi-word terms that can be translated as a unit (usually to a single word) (Levow, Oard, & Resnik, 2005). More complex (i.e., compositional) approaches to phrase translation have also been tried in CLIR systems (e.g., Adriani, 2000; Ballesteros & Croft, 1998; Gao et al., 2001; Monz & Dorr, 2005) but the natural tendency of ranked retrieval systems to reward co-presence of query term translations usually works so well that statistically significant gains over a strong baseline are rarely reported from more sophisticated approaches.

A second possible source of constraints on the translation selection process is syntactic analysis. For example, when part-of-speech information is available in the bilingual dictionary, automatically assigned part-of-speech tags can be used to as a basis for translation selection (Hull & Grefenstette, 1996). Automatically tagging words in short queries with their part of speech can be problematic, however, because short queries offer little context and because the structural cues that are present in queries may be quite different from the structure on which available part-of-speech taggers have been trained. For these reasons, part of speech constraints are better suited to document translation architectures where those problems are less consequential.

A third broad class of techniques for accommodating uncertainty, initially proposed by Pirkola (Pirkola, 1998), exploits the structure induced by the translation process to limit the effect of translation ambiguity. The key idea is to separately estimate the term frequency (TF) and document frequency (DF) of each query term based on the TF and DF of individual translations. More precisely, the estimated TF for a query term in a document is the sum of the TF's for each known translation of that term, while the estimated DF of a query term is the number of documents in that collection that contain at least one known translation for the query term. The DF that results for each query term is lower-bounded by the DF of the most common translation, thereby preventing translations that are rare (and thus highly selective) in the document language from dominating the retrieval results. This approach has since been extended to accommodate translations learned from examples (where any term might conceivably be the translation of any other) by leveraging translation probabilities rather than a limited set of known translations (Darwish & Oard, 2003).

Cases in which no translation is known for a query term have also received considerable attention. When the query and document languages are written using the same character set, it is usually helpful to retain unknown terms with only minor changes (e.g., removal of diacritic marks) in the hope that they will match terms in the document language (as may be the case for proper names and "loan words," for example). When the query and document languages are expressed in different character sets, phonetic transliteration is needed to achieve the same effect (e.g. (Kang & Choi, 2000)).

Another approach to accommodating deficiencies in the translation lexicon is blind relevance feedback, which exploits term co-occurrence to identify additional terms that might plausibly have been included in the query (Ballesteros & Croft, 1997). The basic approach is to mine a collection of documents that is comparable to those that will ultimately be searched, to automatically select some number of top-ranked documents (which have a high likelihood of being on the same topic), and then to automatically select some number of highly discriminating terms that occur more often than chance would predict in those documents. Adding co-occurring terms found in this way to the query before translation can help to overcome gaps in the bilingual dictionary by introducing related terms for which translations are known (McNamee & Mayfield, 2002). Adding co-occurring terms to the query after translation can sometimes achieve a similar effect by augmenting the known translations with related terms that might be unknown (but correct) translations of query terms. The availability of a comparable collection in the document language is rarely a problem (because the collection that is to ultimately be searched can be used), but query-language collections with appropriate characteristics (e.g., genre, topical coverage, and time frame) may be hard to obtain in some cases.

For the experiments reported in this paper we employed Pirkola's structured query method in conjunction with simple dictionaries that specify cross-language synonymy. We chose this option because we expected that searchers would find these types of dictionaries to be familiar, and thus easily understood. Because our experiments involve only European languages, we retained unknown terms (but with diacritic marks removed). We did not employ blind relevance feedback before or after translation because we felt that the additional complexity might hinder the user's development of mental models of system operation.

2.2. Related work on interactive CLIR

Each of the techniques described above was originally developed in the context of fully automatic systems; the active involvement of the user in interactive systems changes the picture considerably. For example, blind relevance feedback is rarely used in interactive monolingual retrieval systems because: (1) it results in system behavior that searchers have difficulty understanding (and therefore controlling), and (2) automatic introduction of inappropriate "related" terms will sometimes adversely affect retrieval effectiveness (and thereby decrease the user's confidence in the system). These factors are likely to be just as important in interactive CLIR applications, and we are aware of no case in which blind relevance feedback has been incorporated in an interactive CLIR system. Structured queries, phrase translation, backoff translation, and phonetic transliteration appear to offer more scope for incorporation into interactive CLIR systems.

One of the first designs for interactive CLIR was the Korea Advanced Institute of Science and Technology (KAIST)/ETRI "FromTo-CLIR" system (Kim et al., 1999). FromTo-CLIR employed fully-automatic query translation to allow searchers to formulate queries in one language that could then be passed to a Web search engine in another. Brief summaries and the full text of retrieved documents were then automatically translated back into the query language for the searchers to review. Fully automated translation can, however, make it hard for the user to understand (and ultimately control) what the machine is doing. When query and document are expressed in the same language, searchers often use their initial queries to discover the terms used by authors to express interesting ideas; those terms can then be used to refine subsequent queries. When examining documents in a language they cannot read, however, cross-language searchers will see terms chosen by

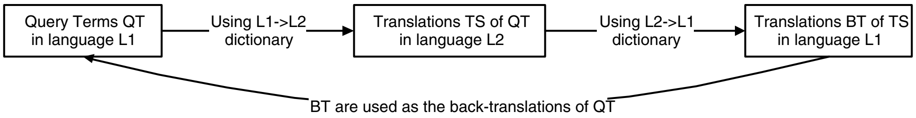

Query Terms QT in language L1
Using L1->L2 dictionary
Translations TS of QT in language L2
Using L2->L1 dictionary
Translations BT of TS in language L1
BT are used as the back-translations of QT

Fig. 3. Back-translations in the MULINEX system.

the translation system, not the terms actually used by authors. If the searcher later includes one or more such terms in a cross-language query, fully automatic query translation may yield document language terms that bear no relation to the searcher's intended meaning, resulting in unexpected and potentially inscrutable results.

The first effort to address that challenge appeared in the New Mexico State University Keizai system (Ogden, Cowie, Davis, & Ludovik, 1999). In Keizai, searchers were required to select appropriate translations after examining English definitions of each translation alternative before the search would be conducted. This "two-stage" process was designed to provide the searcher with greater insight into, and control over, operation of the system, at the cost of some additional effort on the part of the searcher. As a proof of concept, Keizai pointed the way toward system designs that more naturally reflected the task characteristics, but the bilingual dictionaries in which English definitions are available for each translation are relatively rare. Most bilingual dictionaries are designed for use by people with some facility in the target language, and thus they typically present definitions in the same language as the translation.

The German Research Center for Artificial Intelligence's (DFKI) MULINEX system sought to overcome this limitation by dynamically generating lists of potential synonyms for each translation (Capstick et al., 2003). The goal here was to indicate, rather than define, the meaning of the translations that the user might choose. Fig. 3 illustrates the "back translation" strategy that was used in MULINEX. The key insight on which this was based is that synonyms in one language are often translated using the same word in another language. In such cases, reversing the translation process will generate query-language synonyms for the document-language term. These synonyms can then help a monolingual user understand which sense of a homonymous[1] query-language term is represented by a particular translation. Back translation suffers from two limitations, however: (1) when no synonyms can be found in the dictionary, the technique is not helpful; and (2) significant homonymy in the target language can result in an eclectic set of potentially confusing back translations. Our work extends this line of inquiry by augmenting back translation with examples of usage, a complementary source of insight into the meaning of potential translations.

That brief recap describes the state of the art for interactive CLIR at the start of the research reported in this paper. Three related efforts that unfolded concurrently with our work also bear mention here because their results helped to shape our thinking about the design space as our work proceeded. The most fundamental of these was the interactive track of the Cross-Language Evaluation Forum (iCLEF), which brought together researchers working on interactive CLIR each year between 2000 and 2005 to compare results using shared user study designs. The iCLEF evaluations started with a focus on evaluation of document-selection; after 2001, the focus expanded to include end-to-end searching. Participants have included the National Distance Education University (UNED) and the University of Alicante in Spain, the Swedish Institute of Computer Science, the University of Sheffield in the United Kingdom, and our group at the University of Maryland in the United States of America. The user studies reported in this paper were conducted using the 2002 and 2003 iCLEF study designs, and the results of those studies have been previously reported (Dorr et al., 2003; He, Oard, & Plettenberg, 2006; He, Wang, Oard, & Nossal, 2002). This paper draws together those results and includes substantial additional analysis that could not be completed within the time constraints of annual evaluation cycles.

The CLARITY project, a joint effort of the University of Sheffield and five other institutions, was the first reported case in which design of an interactive CLIR system was grounded in a formal assessment of user needs. Using interviews and paper mockups, they found that polyglots (users who know several

[1] In linguistics, polysemy refers to words with different shades of meaning, while homonymy refers to words with unrelated meanings. Homonymy poses the greater challenge for CLIR systems.

languages, often at different levels of proficiency) formed an important user group whose needs had previously been under-studied. That conclusion likely reflects the setting of their work, focusing on educated professionals in a European context. The work we report in this paper has a somewhat different focus; foreign language proficiency is markedly lower in the United States than in Europe, and we are therefore particularly interested in serving users that have little or no reading proficiency in the target language. The CLARITY system initially included the two-stage back-translation design first implemented in MULINEX, but their early user studies indicated some dissatisfaction with the additional effort that was needed before any retrieval results could be examined (Petrelli, Levin, Beaulieu, & Sanderson, 2006). The design of the system was therefore changed to incorporate a progressive refinement strategy in which fully automatic query translation was first performed and then subsequent searches could then be refined by interactive deselection of inappropriate translations. We independently implemented a similar design at about the same time, as described in Section 4.[2]

The UNED iCLEF team explored a direction that is complementary to the work reported in this paper. Building on a well known insight in the machine translation community that multi-word expressions exhibit markedly less homonymy than single words, they generated all possible translations for the constituent words in noun phrases and then filtered the results using a representative text collection to remove all but the most common rendition (Lopez-Ostenero, Gonzalo, Penas, & Verdejo, 2002). Because the resulting noun phrases always induced a single unique translation, user-assisted query translation could be avoided. Instead, query refinement was supported using noun phrases found in highly ranked documents. They found that noun-phrase translation yielded accuracy improvements in time-constrained retrieval tasks when compared with back-translation and that users expressed a preference for noun phrase translation, in part because back-translation sometimes yielded results that were difficult to interpret. Our work took a different direction, seeking to retain the broader range of feedback opportunities offered by user-assisted query translation while augmenting the system-generated feedback about the meaning of alternative translations in order to better support that process.

2.3. Prototype system design

We used these insights as the basis for design of a system to support iterative prototyping that we call the Maryland Interactive Retrieval Advanced Cross-Language Engine (MIRACLE). The system incorporates the following features:

2.3.1. User-assisted query translation

This is designed to foster transparency and control, facilitating the searcher's development of mental models of system operation. Selecting correct translations could improve results, but omitting a useful translation could equally well have an adverse effect. Therefore our principal motivation for including this capability was to support iterative query refinement; if searchers make bad choices, they can see the effect and learn to better control the system. Three types of evidence have been provided to help monolingual searchers determine which translations should be selected: (1) the translation itself, the meaning of which might be recognized by the searcher if it is a loan word or a proper name, (2) a list of possible synonyms (found using back translation), and (3) examples of usage (found in translated or topically-related texts). Because these cues draw information from difference sources, their availability varies. For example, the searchers in our third study (see Section 3.5) issued 259 unique query terms that have total 504 different translations in our dictionary. 363 of these translations have non-trivial back translation (i.e., the back translations contain terms other than the original query term). In the same study, the cues based on parallel text only provided examples of usage for 208 unique translations. Cues based on comparable text achieved slightly higher coverage, but still only provided examples of usage for 283 unique translations.

[2] The precise origin of this idea is difficult to nail down because we had discussed it with the Sheffield team before either team published it.

2.3.2. Rapid adaptation to a new language

Rapid adaptation to a new document language was an important goal in the MIRACLE system design. The query language is always English, so the rich array of language resources that are available for English can be leveraged regardless of the document language. MIRACLE can minimally work with just a simple bilingual term list, but it is designed to readily leverage additional resources when they are available. Results for that aspect of MIRACLE have already been reported, so we do not focus on those points in this paper (He et al., 2003).

2.3.3. Single document language

Integration of result sets from more than one document language can be a useful capability in some applications. For example, in many cases, searchers will want to see relevant documents in the query language when such documents are available. We did not include such a capability in MIRACLE, however, because doing so would have confounded the design of user studies in which our focus was on cross-language searching. Research on integration of search results did proceed concurrently with our work at other sites (e.g. (Braschler, 2004)), although we are not yet aware of user studies that explore the utility of such a capability in interactive settings.

The top part of Fig. 4 shows a typical data flow for a CLIR system based on fully automatic query translation, while the bottom part of that figure shows the data flow for the MIRACLE system. Fully automatic query translation affords the searcher with just one refinement opportunity: reformulation of the query based on examination of search results. In the MIRACLE data flow, by contrast, four refinement opportunities exist. The centerline (forward) path yields the initial search results, including one refinement opportunity (translation deselection based on evidence about the meanings of available translations in the user-assisted query translation component) The backward branches in the MIRACLE data flow illustrate three additional refinement opportunities: (1) query reformulation based on evidence about the meanings of available translations, (2) query reformulation based on examination of search results, and (3) translation deselection and/or reselection based on examination of search results. The key idea is to leverage the speed, memory, scalability and consistency of the machine to provide translation alternatives, while leveraging human reasoning and pattern recognition abilities to control the machine's behavior. For searchers with some knowledge of the document language, we could think of this process as the system helping the searcher translate their query. But it is more natural to think about the searcher helping the system when the searcher does not have document-language skills. That is why we call the approach ‘‘user-assisted query translation’’.

We built MIRACLE using a Java client-server architecture in order to balance easy integration of component technologies (on the server side) with rich interaction in a portable framework (on the client side). Extensive logging functions were provided on the server side to support use of the system for user studies. Our primary goal for this version of MIRACLE was to evaluate interaction strategies, so processing was done offline whenever possible in order to minimize the need for a focus on run-time efficiency at this early stage in our development process.

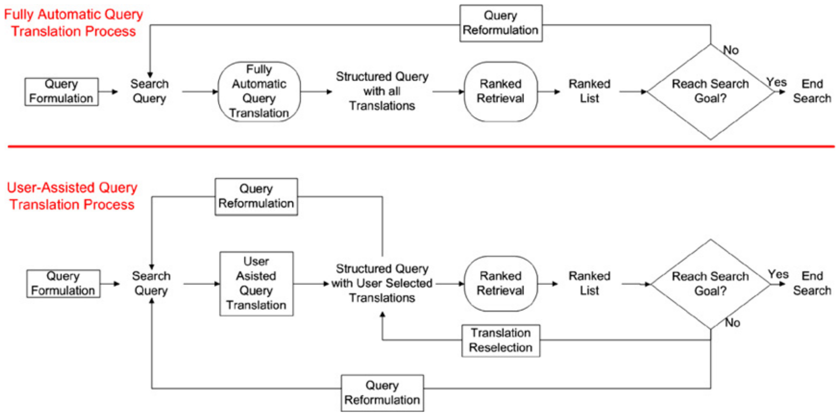


Fig. 4. Data flow for fully automatic query translation (top) and user-assisted query translation (bottom). Rounded corners indicate automatic processes.

2.4. Support for interaction

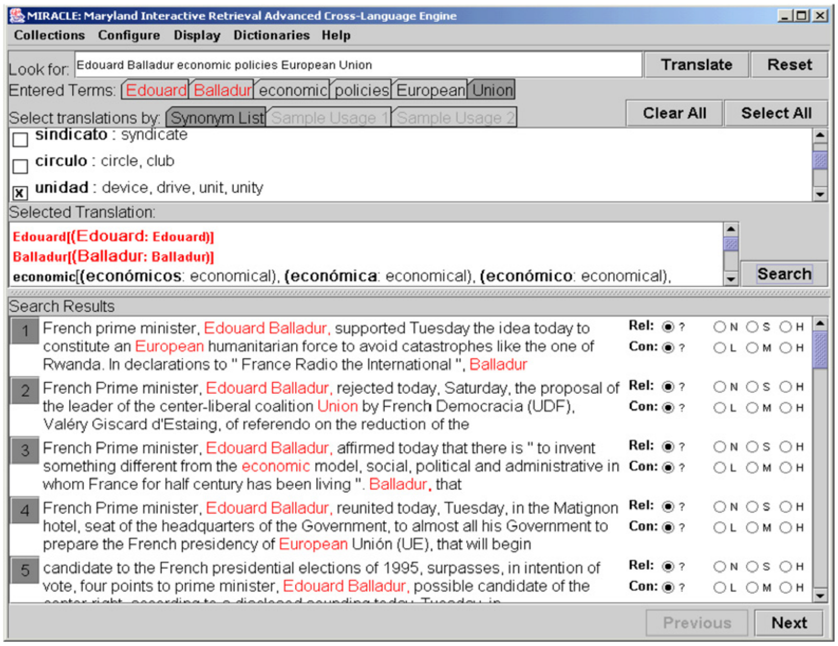
The user's understanding of MIRACLE's capabilities is shaped by the user interface. Our interface design was guided by two key design guidelines: (1) expose interaction opportunities to the user in a straightforward and easily understood manner, and (2) provide immediate feedback in response to control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

As shown in Fig. 5, the MIRACLE interface consists of five major components: *query input*, *translation selection*, *translated query display*, *ranked list display*, and (in Fig. 7) *document display*. Searchers type their queries into a text box, just as they would in a monolingual Web search engine. At present, only unstructured ("bag of words") queries are supported. English stopwords are removed prior to query translation.

When the searcher clicks the "Translate" button, the system obtains all translations for each (non-stopword) query term from the translation lexicon and makes them available for display in the translation selection area. The translation selection area allows the user to choose a query term to work on (using the top set of tabs) and then to select or deselect translations for that term. This function is available only for terms with two or more translation alternatives, and among those the query term with the fewest translation alternatives is initially selected by the system. Translations are presented with different types of cues that searchers can use as a basis for selecting or deselecting translations. The searcher can cycle through alternative types of cues by clicking the tabs above the display area.

As the user selects and deselects translations, those changes are reflected in the next area down, in which the full translated query is continuously displayed. Because the searcher cannot be expected to read document-language terms, the translations are grouped (one per line) by query term, with one back-translation shown for each selected translation of that query term. The interface can be configured to permanently hide those two areas to create contrastive conditions in which the query translation process is always fully automatic (in that case, the "Translate" and "Search" functions are combined, and the button is labeled "Search"). Fig. 6 shows the interface in that condition.

When the user selects "Search," the results are displayed as brief summaries that are sorted in order of decreasing system-assigned score (which, hopefully, reflects a decreasing probability of relevance). Because



Look for: Edouard Balladur economic policies European Union

Entered Terms: Edouard Balladur economic policies European Union

Select translations by: Synonym List Sample Usage 1 Sample Usage 2

□ **sindicato** : syndicate

□ **circulo** : circle, club

☒ **unidad** : device, drive, unit, unity

Selected Translation:

Edouard[(Edouard: Edouard)]
Balladur[(Balladur: Balladur)]
economic[(**económicos**: economical), (**económica**: economical), (**económico**: economical),

1 French prime minister, Edouard Balladur, supported Tuesday the idea today to constitute an European humanitarian force to avoid catastrophes like the one of Rwanda. In declarations to " France Radio the International ", Balladur — **Rel:** ● ? ○ N ○ S ○ H **Con:** ● ? ○ L ○ M ○ H

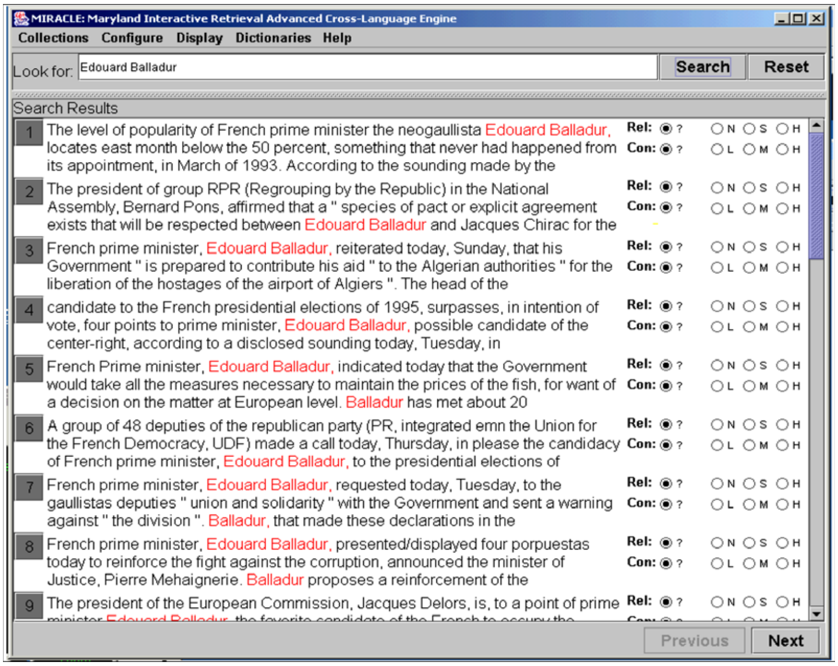
2 French Prime minister, Edouard Balladur, rejected today, Saturday, the proposal of the leader of the center-liberal coalition Union by French Democracia (UDF), Valéry Giscard d'Estaing, of referendo on the reduction of the — **Rel:** ● ? ○ N ○ S ○ H **Con:** ● ? ○ L ○ M ○ H

3 French Prime minister, Edouard Balladur, affirmed today that there is " to invent something different from the economic model, social, political and administrative in whom France for half century has been living ". Balladur, that — **Rel:** ● ? ○ N ○ S ○ H **Con:** ● ? ○ L ○ M ○ H

4 French Prime minister, Edouard Balladur, reunited today, Tuesday, in the Matignon hotel, seat of the headquarters of the Government, to almost all his Government to prepare the French presidency of European Unión (UE), that will begin — **Rel:** ● ? ○ N ○ S ○ H **Con:** ● ? ○ L ○ M ○ H

5 candidate to the French presidential elections of 1995, surpasses, in intention of vote, four points to prime minister, Edouard Balladur, possible candidate of the — **Rel:** ● ? ○ N ○ S ○ H **Con:** ● ? ○ L ○ M ○ H

Fig. 5. The MIRACLE CLIR system, configured for Spanish. The radio buttons to the right of each summary allow recording degree of relevance (Not, Somewhat, Highly) and confidence in that judgment (Low, Medium, High).


MIRACLE: Maryland Interactive Retrieval Advanced Cross-Language Engine
Collections
Configure
Display
Dictionaries
Help
Search
Reset
Previous
Next

Look for: Edouard Balladur

Search Results

1 The level of popularity of French prime minister the neogaullista Edouard Balladur, locates east month below the 50 percent, something that never had happened from its appointment, in March of 1993. According to the sounding made by the
Rel: ● ? ○ N ○ S ○ H
Con: ● ? ○ L ○ M ○ H

2 The president of group RPR (Regrouping by the Republic) in the National Assembly, Bernard Pons, affirmed that a " species of pact or explicit agreement exists that will be respected between Edouard Balladur and Jacques Chirac for the
Rel: ● ? ○ N ○ S ○ H
Con: ● ? ○ L ○ M ○ H

3 French prime minister, Edouard Balladur, reiterated today, Sunday, that his Government " is prepared to contribute his aid " to the Algerian authorities " for the liberation of the hostages of the airport of Algiers ". The head of the
Rel: ● ? ○ N ○ S ○ H
Con: ● ? ○ L ○ M ○ H

4 candidate to the French presidential elections of 1995, surpasses, in intention of vote, four points to prime minister, Edouard Balladur, possible candidate of the center-right, according to a disclosed sounding today, Tuesday, in
Rel: ● ? ○ N ○ S ○ H
Con: ● ? ○ L ○ M ○ H

5 French Prime minister, Edouard Balladur, indicated today that the Government would take all the measures necessary to maintain the prices of the fish, for want of a decision on the matter at European level. Balladur has met about 20
Rel: ● ? ○ N ○ S ○ H
Con: ● ? ○ L ○ M ○ H

6 A group of 48 deputies of the republican party (PR, integrated emn the Union for the French Democracy, UDF) made a call today, Thursday, in please the candidacy of French prime minister, Edouard Balladur, to the presidential elections of
Rel: ● ? ○ N ○ S ○ H
Con: ● ? ○ L ○ M ○ H

7 French prime minister, Edouard Balladur, requested today, Tuesday, to the gaullistas deputies " union and solidarity " with the Government and sent a warning against " the division ". Balladur, that made these declarations in the
Rel: ● ? ○ N ○ S ○ H
Con: ● ? ○ L ○ M ○ H

8 French prime minister, Edouard Balladur, presented/displayed four porpuestas today to reinforce the fight against the corruption, announced the minister of Justice, Pierre Mehaignerie. Balladur proposes a reinforcement of the
Rel: ● ? ○ N ○ S ○ H
Con: ● ? ○ L ○ M ○ H

9 The president of the European Commission, Jacques Delors, is, to a point of prime minister Edouard Balladur, the favorite candidate of the French to occupy the
Rel: ● ? ○ N ○ S ○ H

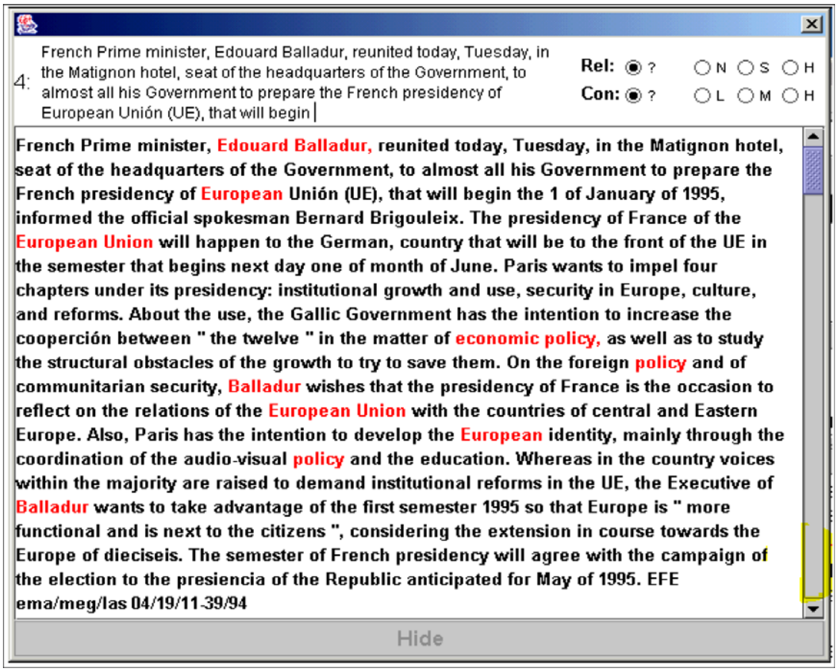
Fig. 6. The MIRACLE system with user-assisted query translation disabled.

we assume that the searcher has little or no reading ability in the document language, both summaries and full documents are presented as English translations. In prior work, we compared two options: (1) using word by word translation, which we called “gloss translation;” or (2) using a machine translation system to translate the documents. Our results in that case indicated that both approaches were viable, but that machine translation resulted in higher accuracy and less fatigue (Wang & Oard, 2001). Both approaches are supported in MIRACLE, but machine translation was used for all of the studies reported in this paper. Ten summaries are displayed per page, although the user may need to scroll to see all 10. A total of 10 result pages are available (using the “Next” and “Previous” buttons), so the searcher can examine as many as one hundred documents. Dividing the result set in this way facilitates rapid delivery of search results when network bandwidth is limited. All terms that share a common stem with any non-stopword query terms are highlighted (in red)[3] to draw the searcher’s eye.

Each summary is labeled with a numeric rank (1, 2, 3, ...) that is displayed on a button to its left. The full text of any document can be viewed in a pop-up window by clicking on the appropriate button. In order to maintain context, the numeric rank of the document and the text of the summary are shown at the top of the document examination window. Fig. 7 illustrates a document examination window.

MIRACLE can be configured to record explicit relevance judgments made by the searcher, a useful capability for some user study designs. Three degrees of relevance can be indicated (Not relevant: “N,” Somewhat relevant: “S,” and Highly relevant: “H”). A fourth value, “?” (indicating Unjudged), is initially selected by the system. Similarly, the searcher can optionally indicate their degree of confidence in their judgment as (Low: “L,” Medium: “M,” or High: “H”), with a fourth value (“?”) being initially selected by the system. Searchers can record relevance judgments and confidence values in either the ranked list of summaries or in a pop-up document examination window (when that window is displayed). MIRACLE logs the times at which documents are selected for examination and the times at which relevance judgments for those documents are recorded. That data allows computation of the approximate examination duration for each document.

[3] For interpretation of the references in color in this figure, the reader is referred to the Web version of this article.



4: French Prime minister, Edouard Balladur, reunited today, Tuesday, in the Matignon hotel, seat of the headquarters of the Government, to almost all his Government to prepare the French presidency of European Unión (UE), that will begin

Rel: ● ? ○ N ○ S ○ H

Con: ● ? ○ L ○ M ○ H

French Prime minister, Edouard Balladur, reunited today, Tuesday, in the Matignon hotel, seat of the headquarters of the Government, to almost all his Government to prepare the French presidency of European Unión (UE), that will begin the 1 of January of 1995, informed the official spokesman Bernard Brigouleix. The presidency of France of the European Union will happen to the German, country that will be to the front of the UE in the semester that begins next day one of month of June. Paris wants to impel four chapters under its presidency: institutional growth and use, security in Europe, culture, and reforms. About the use, the Gallic Government has the intention to increase the coopercíón between " the twelve " in the matter of economic policy, as well as to study the structural obstacles of the growth to try to save them. On the foreign policy and of communitarian security, Balladur wishes that the presidency of France is the occasion to reflect on the relations of the European Union with the countries of central and Eastern Europe. Also, Paris has the intention to develop the European identity, mainly through the coordination of the audio-visual policy and the education. Whereas in the country voices within the majority are raised to demand institutional reforms in the UE, the Executive of Balladur wants to take advantage of the first semester 1995 so that Europe is " more functional and is next to the citizens ", considering the extension in course towards the Europe of dieciseis. The semester of French presidency will agree with the campaign of the election to the presiencia of the Republic anticipated for May of 1995. EFE ema/meg/las 04/19/11-39/94

Fig. 7. The document examination pop-up window.

2.5. Performing the search

MIRACLE communicates with the search system through the file system in order to facilitate integration of newly developed search capabilities. The initial implementation was based on the InQuery text retrieval engine (version 3.1p1) from the University of Massachusetts. InQuery's "synonym" operator provides native support for Pirkola's structured query technique, and it is therefore a suitable system when translation probabilities are not available (Pirkola, 1998). When the client provides a query, the server informs the client of the known translations for each query term. The searcher can then optionally elect to deselect some translations. When the user clicks "Search," the client informs the server which translations remain selected. The server then formulates a structured query containing only the selected translations using InQuery syntax and stores that query in a file. The server then initiates an InQuery search, which reads the query from the file and writes a ranked list in the standard format used for evaluation in the Text Retrieval Conferences (TREC). When the search completes, the server parses the TREC-format results file, creates query-focused summaries for the top 10 search results, and passes them to the client for display. The summaries are generated by searching through the content part of the documents and identifying up to three sentences that contain query terms. InQuery provides a rich Application Programmer Interface (API) that could have been used to achieve tighter integration, but by adopting a file-passing process we substantially simplified the integration of alternative search systems.

When no translation is known for the exact form of a query term that is provided by the searcher, MIRACLE automatically tries a backoff translation strategy (Resnik et al., 2001). This occurs regardless of which search system is used. The server stores two hash tables, one keyed to the terms as they appear in the dictionary and one keyed to their stems. If no translation was found in the first hash table for the exact ("surface") form of an English query term, that term is stemmed (using the Porter stemmer) and tried again using the same hash table. If that fails, the second hash table is consulted using the stemmed form of the query term. If that still fails, the untranslated term is retained in the query (in the hope that it might be a proper name or a loan word that will match) and the client is informed that the term is untranslatable so that feedback can be provided to the searcher. The searcher might then choose to replace that query term with a near-synonym for which translations might be known. The backoff process stops when the first match is found, thus minimizing the introduction of spurious translations.

As Fig. 2 depicts, the user may iterate their search process at any point, returning from query translation to query reformulation without searching, returning from examination of search results or an individual document to reselect alternative translations, or returning from examination of search results or an individual document to reformulate their query. The first three of these five options are unique to the user-assisted query translation process, and thus of particular interest in our user studies.

3. User studies

Ultimately, we want to understand the degree to which the systems we build will help real searchers to accomplish real tasks. Fundamentally, there are three ways in which we might try to learn this. The most widely used study design is to model the users task in a manner that permits automated evaluation. This is the approach adopted in TREC, for example: the user is modeled as posing a query and wishing to receive a ranked list of documents with topically relevant documents near the top of that list. Fully automated study designs with reusable evaluation resources have also been defined for summarization and machine translation. It is therefore possible to separately evaluate many of the components that are needed for interactive CLIR. Modeling interaction is considerably more difficult, however, since the complex interplay between perception and cognition resists the type of binary (right/wrong) characterization that has been so successfully employed in automated evaluation frameworks for information retrieval, summarization, and machine translation.

At the other end of the spectrum, we can learn a lot by building real systems, giving them to real searchers, and watching them accomplish real tasks. Observational studies of that type are often employed when mature system designs and experienced searchers are available. In such cases, rich description and systematic qualitative analysis methods can offer insights into interactions between system, searcher, and task that would difficult to capture using more highly structured study designs. Because the data collection and analysis for observational studies can be expensive, controlled user studies in which two systems are compared quantitatively are often used when the research questions to be explored can be crafted in a manner that is sufficiently narrow. Because both approaches bring strengths and weaknesses, mixed-method studies that wrap richer data collection around a quantitative controlled study design can often affordably yield deeper insights than would be the case were either approach to be tried in isolation. We therefore adopted a mixed-methods study design for the experiments reported below.

Our user studies were designed to explore the following research questions:

- Can cross-language searchers find documents more effectively when we give them some degree of insight into and control over the query translation process?
- How do cross-language searchers adapt their search process when user-assisted query translation is available?
- Do cross-language searchers prefer to exercise control over the query translation process? What reasons do they give for their preference?

We conducted three experiments (two in April 2002, and one in April 2003) using variants of the same study design to focus on different aspects of the interaction. We first present the common settings of the design that were shared by the three experiments, then explain how each experiment differed.

3.1. Within-subjects controlled user study design

We refer to a user study design as “controlled” if we assign the tasks, systems, and order of completion rather than allowing the searchers to make those choices for themselves. Controlled study designs require that we sacrifice some fidelity (e.g., by basing each search on assigned rather than internalized information needs) in exchange for the potential to aggregate similar conditions during analysis. In a within-subjects design, each subject (i.e., searcher) performs repeated trials (e.g., several searches, each for a different topic). The order of those trials is varied systematically in order to block (i.e., average out) the effects of presentation order on learning and fatigue and the effects of individual differences in searchers and topics on topic difficulty, thereby focusing on the desired effect (the different systems). A Latin square (a square matrix in which no two rows or columns contain the same sequence of conditions) was used as the basis for structuring presentation order. Because within-subjects designs require repeated trials, they require a significant time investment from each searcher. They are, however, generally more economical than the alternative between-subjects designs in which only population averages can be compared, both because within-subjects designs offer more scope for characterizing individual differences and because the time required to recruit subjects and train them to use the systems can be amortized over multiple trials (Maxwell, Dalaney, & Dimmick, 2003).

In our study design, the independent variable (the effect of which we wish to study) is the system design. We compared two conditions, the *user-assisted* system (using the full capabilities of MIRACLE) and the *automatic* system (the same system, but with the query translation and translated query display areas permanently hidden). As a dependent variable, we chose the F measure, a widely reported measure of the degree to which a set contains all and only documents that are topically relevant. Formally, F is a weighted harmonic mean of recall and precision, parameterized by a factor β that characterizes the relative importance of precision. We chose $\beta = 2.0$, which makes F more sensitive to improvements in precision (the fraction of documents found by the searcher that are actually relevant) than to improvements in recall (the fraction of available relevant documents that are found by the searcher). We favored precision because of the widely reported tendency of users in many situations to satisfice, stopping their search when adequate information is available rather than continuing until all potentially useful information has been found (Simon et al., 1986).

We chose a measure based on topical relevance as a dependent variable because topicality has been shown to exhibit a useful degree of consistency across assessors (Voorhees, 2000). Topical relevance is encoded as a binary (yes/no) variable in many information retrieval experiments, but it is sometimes perceived by searchers as a matter of degree. We therefore asked our searchers to indicate three levels of relevance (none, somewhat, or highly) and to also indicate their degree of confidence in their judgment. We compute the F measure twice, once with *strict* judgments from the searcher (only judgments of highly relevant for which at least moderate confidence was reported) and once with *loose* judgments (judgments of somewhat or highly relevant, regardless of the reported confidence). Strict judgments seek to minimize the confounding effect of differing opinions of relevance, sometimes at the expense of data sparseness, while loose judgments strike the opposite balance.

To gain additional insights into the nature of the user's activity, we computed the total number of query iterations, especially those iterations involved translation selection and deselection. We augmented this query-oriented measure with a qualitative study in which we examine how query terms were generated, how often certain search strategies and tactics were used, and which factors that could affect the performance of the user-assisted translation selection method.

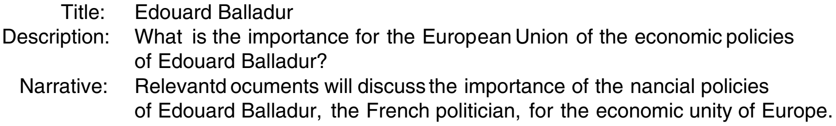
After an initial training session, the participants were given a fixed length of time for each search session to identify relevant documents. They were asked to emphasize precision over recall. Specifically, searchers were told that "more credit will be awarded for accurately selecting relevant documents than for the number of documents that are selected, because in a real application you might have to pay to have a high-quality translation prepared for each selected document." We asked each participant to fill out brief questionnaires before the first search session (for demographic data), after each topic, and after using each system. Each participant used the same system at a different time, so we were able to observe each individually and make extensive observational notes. We also conducted a semi-structured interview (in which we tailored our questions based on our observations) after all searches were completed.

3.2. Evaluation resources

Computing the F measure requires that we have a set of topics, a set of documents, and a set of relevance judgments for every topic-document pair. We obtained these resources from the Cross-Language Evaluation Forum (CLEF).[4]

Because the query language of MIRACLE is English, we chose document languages other than English. In 2002, we elected to work with the CLEF German document collection, which contained 71,677 news stories

[4] Additional details on resources used in the 2002 and 2003 CLEF interactive track evaluations are available at http://nlp.uned.es/iCLEF/.



Title: Edouard Balladur
Description: What is the importance for the European Union of the economic policies of Edouard Balladur?
Narrative: Relevantd ocuments will discussthe importance of the financial policies of Edouard Balladur, the French politician, for the economic unity of Europe.

Fig. 8. A topic statement.

from the Swiss News Agency (SDA) and 13,979 news stories from Der-Spiegel. For our 2003 user studies, 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 searcher). The Systran translation system is fast enough to translate individual documents on demand (at about one second per document). Our decision to translate all of the documents in advance was made solely to simplify the implementation of our prototype system. All searches were performed using dictionary-based query translation rather than searching the documents that had been automatically translated into English.[5]

CLEF topics are initially proposed in written form by CLEF relevance assessors based on their own interests and their understanding of the topical coverage of the available document collections. For each topic, a written topic statement is created and vetted by other assessors to ensure that individuals other than the creator of the topic can clearly determine whether specific documents would be relevant. Fig. 8 shows an example of a topic statement. The title field is typically rendered in the keyword-oriented telegraphic style that is typical of Web queries. The description field, usually used in conjunction with the title field, is intended to represent what a searcher might initially say to someone who was helping them with their search. The narrative field, together with the other two fields, is intended to provide additional information that may be needed to make accurate relevance judgments. In our user studies, we showed all three fields to our searchers because we wanted them to approximate the judgments made by CLEF relevance assessors to the greatest degree possible.

The first stage in the CLEF relevance assessment process is translation of the topics into the language of the documents. This is done manually, typically by the person who will ultimately perform the relevance assessment. Translations are reviewed at this stage in order to resolve any differences in interpretation. For each topic, highly ranked documents are obtained from several fully automatic information retrieval systems, and each such document is judged for relevance to the topic using the translated topic statement. These judgments are prepared by native speakers of the language in which the documents are written. Questions of interpretation can be resolved at this point through discussion among the assessors. This "pooled relevance assessment process" yields an initial set of relevance judgments that provides the basis for topic selection in the iCLEF experiments.

CLEF produces 40 topics each year, far more than any one searcher could hope to complete during a user study. The iCLEF user studies therefore used small subsets of these topics, four in 2002 and eight in 2003. In earlier work, we had learned that "broad" topics for which relevant documents addressing many aspects of the same event could be found in the collection (e.g., "*Conference on birth control*") yielded results that were difficult to compare with "narrow" topics for which the available reporting addressed just a single aspect of the topic (e.g., "*Bush fire near Sydney*") (Oard et al., 2004). We therefore focused on broad multi-aspect topics for our experiments. Among the available multi-aspect topics, those with a moderate number of known relevant documents that the iCLEF organizers felt could be reliably assessed without specialized knowledge were preferred. Table 1 lists the topics used in our experiments.

Voorhees has shown that relevance judgments created by assessing only highly ranked documents from automatic systems can be reliably reused to evaluate other automatic systems, but that interactive searchers

[5] Although it is certainly possible to translate a collection of this size (on a single machine, in a few days), such an approach does not scale well. For example, translating the entire English Web into every query language supported by modern Web search engines would require centuries of machine time. Term-oriented techniques can be efficiently applied to document translation fairly easily (e.g., (Oard & Ertunc, 2002)) but search architectures based on document translation lack the opportunity for productively involving the user that user-assisted query translation provides.

Table 1
The titles of the 12 topics used in our experiments

Topic ID	Topic title	Number of relevant documents
1	Genes and Diseases	22
2	Treasure Hunting	47
3	European Campaigns against Racism	25
4	Hunger Strikes	68
5	The Ames espionage case	50
6	European car industry	36
7	Computer Security	34
8	Computer Animation and Film	3
9	Edouard Balladur	34
10	Marriage Jackson-Presley	30
11	German Armed Forces Out-of-area	23
12	EU fishing quotas	181

often find additional relevant documents that no automated system discovered (Voorhees, 2000). The iCLEF relevance assessment process therefore included a second stage of relevance assessment for every previously unassessed document for which a judgment was recorded by any participant in an iCLEF user study. This second-stage assessment process was performed a year after the first stage assessments were completed using the same process employed for the first stage. Often, it was done with the same assessor. The result was a rich set of relevance judgments in which careful "ground truth" assessments by a native speaker are available for every document seen by any user. This is sufficient to completely characterize precision, and to compute relative recall (i.e., recall relative to the set of known relevant documents, rather than to the set of all existing relevant documents).

3.3. Language resources

Language resources are not typically distributed with a test collection, but for a CLIR experiment the available language resources can be as important a factor for defining the conditions of an experiment as the topics, documents, and relevance judgments. We obtained a German-English bilingual term list from the Chemnitz University of Technology,[6] which provides translations for 102,402 unique English words. We used the German stemmer from the "Snowball" project to stem both the German collection and the German translations of the query terms.[7] No decompounding was performed. Our Spanish-English bilingual term list, which contains 24,278 English words, was constructed from multiple sources (Habash, 2003). We used InQuery's Spanish stemmer to stem both the collection and the Spanish translations of the English queries. In other studies, we have found that dictionaries of this size yield average measures of retrieval effectiveness that are near the limit of what can be achieved using dictionary-based techniques; dictionaries tend to grow by adding progressively less common words, and eventually the words become so uncommon that they rarely occur in queries (Demner-Fushman & Oard, 2003).

One of our techniques for generating examples of usage requires parallel (i.e., translation-equivalent) text; we obtained that from the Foreign Broadcast Information Service (FBIS) TIDES data disk, release 2. Our other technique requires a large English monolingual collection with documents from a similar genre. We had two choices at the time of the experiment. We could have used the CLEF English corpus, which shares a common time frame with the Spanish collection being used in the experiment. We chose instead to use the TDT-4 collection, which is from a period about 6 years after that of the CLEF Spanish collection. This choice allowed us to minimize the chance that some examples of usage might come from relevant query-language documents, thus allowing us to focus on the utility of the technique for identifying representative examples of usage without additional confounding factors.[8] Of course, in an operational setting the fact that some

[6] http://dict.tu-chemnitz.de/.
[7] http://www.snowball.tartarus.org/.
[8] The TDT-4 collection is available from the Linguistic Data Consortium (http://www.ldc.upenn.edu).

examples of usage might be generated from relevant documents would be an additional benefit that should not be artificially suppressed.

3.4. Participant profiles

Among the three experiments, we recruited a total of 20 participants. The participant population was relatively homogeneous across the three experiments:

Native English speakers with limited 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 reported good reading skills in the document language (German, participating in Study 1).

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. 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 searched at least once or twice 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 seven male participants.

Not previous study participants
None of the participants had previously participated in a TREC or iCLEF study.

3.5. Three user studies

The principal goal of our first study was to measure the effect of user-assisted query translation on retrieval effectiveness and to learn about the subjective views of our study participants. This served as a baseline against which results from the other two studies could be compared. In order to maximize the opportunity to explore user-assisted query translation, we used relatively long search sessions (20 min for each topic). This decision resulted in limiting the number of topics that each user could work with; a 2.5 hour period provided adequate time for four topics (topics 1–4 in Table 1), including training, breaks, questionnaires, and interviews. We recruited four participants for this study, and German was the document language.

The study we refer to in this paper as "Study 2" was actually conducted before the study described above, but it is most naturally thought of as a contrastive condition. This second study used an identical experiment design to the first one, with two important exceptions: (1) there were eight participants rather than four, and (2) fewer relevant documents were available for the same set of topics. The CLEF German collection includes

Table 2
Number of relevant documents by topic in each part of the CLEF German collection

	Number of relevant documents		
	Der Spiegel (Study 2)	SDA	Total (Study 1)
Topic 1	12	10	22
Topic 2	21	26	47
Topic 3	0	25	25
Topic 4	12	56	68

documents from Der Spiegel and SDA, but for this second study we indexed only the Der-Spiegel collection, which contains 13,799 German news articles. In Study 1, the average number of relevant documents for a topic was 47 (range 22–68), while in Study 2 that average was 11 (range 0–21), as shown in Table 2. This change was actually a mistake; Study 2 had originally been intended as our iCLEF 2002 experiment. By unintentionally indexing only part of the document collection we serendipitously gained the ability to examine the effects of the number of relevant documents on the user-assisted query translation process. Of particular note is the case of Topic 3, for which no relevant documents are known in the Der Spiegel collection. Topics that lack relevant documents are not normally used in information retrieval evaluations because they would be of no value for characterizing differences in retrieval effectiveness using recall and precision measures. Unproductive topics do occur in real applications, however, so including such cases in studies of interactive systems can yield useful insights.

Our third study was designed to deepen our understanding of search behaviors when user-assisted query translation was available. Compared to the first two studies, the design of this experiment had three major changes: (1) Spanish was chosen as the document language, (2) the number of topics searched by each participant was increased to 8 (topics 5–12 in Table 1), and (3) the search session for each topic was reduced to 10 min in order to avoid requiring more than a half day from each participant. Studies of (monolingual) Web searching show that search sessions are often short (e.g., (Spink & Jansen, 2004)), so characterizing the effect of task duration on the utility of user-assisted query translation is desirable. As with Study 2, we recruited 8 participants for this third study. The only significant difference in the MIRACLE system implementation for Study 3 was that we provided examples of usage from (same-genre) comparable text rather than from (translation-equivalent) parallel text. Only examples drawn from parallel text had been used in Studies 1 and 2.

In all three studies, each participant performed the task individually in the presence of an observer who was familiar with the goals of the study. The observer made notes during each session to record their impressions, and then used those notes as a basis for focusing the discussion in a semi-structured interview once all searches were completed. The observer's notes from the search sessions and the interview were available along with log files generated by MIRACLE and questionnaires completed by the participants for analysis. We enhanced our logging ability by incorporating screen capture (using Camtasia Studio) and by recording comments made during the search sessions and the subsequent semi-structured interview on audio tape. Participants in all three studies were encouraged to comment at any time on points that they wished to have noted by the observer.

3.6. Retrieval effectiveness results

Fig. 9 depicts retrieval effectiveness on a topic-by-topic basis for each of the three studies. Each vertical bar on the left side of the bar graphs for Study 1 represents the average of $F_{\beta=2}$ across two participants; bars on the left side in Studies 2 and 3 are averages across 4 participants. In each case, the rightmost bars represent averages over all topics (averaged over 8 trials each for Study 1, 16 each for study 2, and 32 trials each for Study 3). Positive effects on retrieval effectiveness from user-assisted query translation are evident (on average) for Study 1 (20-min sessions, German) and Study 3 (10-min sessions, Spanish), but Study 2 shows the opposite effect. In Study 1 user-assisted query translation yielded a 48% relative increase in $F_{\beta=2}$ over the automatic condition (0.4995 vs. 0.3371) with strict relevance judgments, although that difference was not found to be statistically

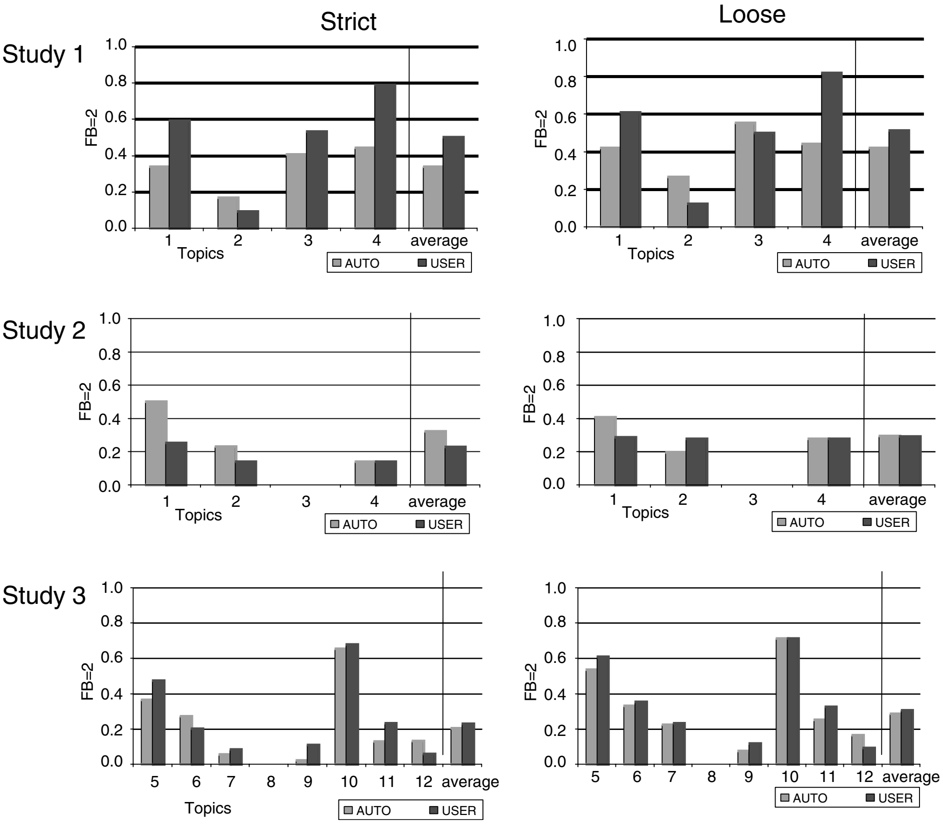


Fig. 9. Retrieval effectiveness ($F_{\beta=2}$) for the user-assisted and the automatic conditions over three user studies. Documents judged "partially relevant" are treated as relevant in the loose condition and as non-relevant in the strict condition.

significant by a two-factor ANOVA with replication ($p = 0.11$).[9] Measurement using loose judgments yielded a smaller (22%) increase in $F_{\beta=2}$ (0.5095 vs. 0.4176).

Comparing the average $F_{\beta=2}$ values of Study 2 with those in Study 1 suggests that the number of available relevant document can have a substantial effect on retrieval effectiveness. Retrieval effectiveness declined more markedly for the three topics that participants in Study 1 had gotten good results with than for the one topic with which they had difficulty in Study 1 (topic 2). The drop in retrieval effectiveness was largest for the user-assisted condition, which had yielded the best results in study 1. Studies 1 and 2 had a disjoint set of participants, however, so it is possible that some portion of the observed difference results from individual differences that can not be controlled for in this comparison. Comparing within Study 2, evaluation with strict relevance judgments yielded an apparent (but not statistically significant) decrease in $F_{\beta=2}$ from the automatic condition to the user-assisted condition (0.3206 vs. 0.2268). Evaluation using loose relevance judgments yielded no measurable difference in retrieval effectiveness (0.2889 vs. 0.2931).

Making sense of these results requires considering the task that we set for our participants. Our automatic condition employed Pirkola's method, which can be thought of as recall-oriented. Although relevant documents may appear lower in the ranked list than they would with careful selection of appropriate translations, Pirkola's method ensures that every possible translation of each query term makes at least some contribution to the final score assigned to each document. Thus, a determined user with sufficient time has a chance of finding relevant documents even if they appear lower in the ranked list. In the user-assisted

[9] The sample sizes in these experiments are small enough that statistically significance would be unlikely even when true differences are present.

condition, by contrast, there is some risk that the searcher might completely eliminate an appropriate translation and therefore prevent the inclusion of some relevant documents in the ranked list. When relevant documents are plentiful and the vocabulary they use is diverse (e.g., because they come from different sources), this may not pose a serious problem for a precision-focused searcher. But when only a few relevant documents are available from a single source, infelicitous deselection of an appropriate translation could limit the opportunity to find those documents. Additional studies would be needed to confirm or refute this conjecture.

Briefer searches seemed to benefit less from user-assisted query translation than longer ones. In Study 3 (with 10-min searches) we observed only a small (and not statistically significant) 13% apparent increase in averaged $F_{\beta=2}$ for the user-assisted condition with strict relevance judgments (0.2272 vs. 0.2014), and similar results were obtained for loose judgments. Topic-by-topic comparison with the previous studies is not possible in this case because different topics (and documents) were used. Moreover, the searchers themselves cannot help to explain the difference because different searchers participated in each study. One plausible explanation is that searchers may typically initially explore the document collection by searching with a variety of queries, thus deferring the use of user-assisted query translation to later in the process. An analysis of log files from Studies 1 and 2 supports this conjecture; in total, searchers used the translation selection panel to change their searches 34 times during the first 10 min and 55 times during the remaining 10 min in those 20-min sessions.

It is notable that one topic in Study 3 (topic 8) yielded no relevant documents for any searcher, this was likely because there were only three known relevant documents for topic 8 (see Table 1). More generally, as Fig. 9 clearly indicates, retrieval effectiveness exhibited substantial variation across topics. This is not at all unexpected; similar effects are commonly seen in both interactive and fully automatic evaluations of retrieval effectiveness(e.g., (Voorhees, 2000)). This has some important consequences for analyses based on overall averages. Consider, for example, Topic 9 with strict judgments. In that case, user-assisted query translation yielded a 500% improvement in $F_{\beta=2}$ (0.1079 vs. 0.0202). The absolute improvement for topic 5 was similar (0.4729 vs. 0.3619), but the relative improvement was far smaller, only 8%. Averaging across those two cases results in reporting a 48% relative improvement (0.2904 vs. 0.1911). The upshot of this is that averages computed in the usual way tend to understate differences that are observed on difficult topics—exactly the topics for which we might hope user-assisted query translation would have the greatest scope to be helpful. If we focus only on the two cases in which the automatic condition yielded $F_{\beta=2} < 0.1$ using strict relevance judgments, we see that user-assisted query translation apparently helped in both cases. While two cases are too few to reliably support any broader inference (and the results would be more mixed if we were to choose $F_{\beta=2} < 0.2$ as our cutoff), this way of looking at the results does suggest that in future studies it might be productive to focus on topics that it is expected users would find difficult. Doing so would require that we develop some way of predicting topic difficulty in an interactive setting. Perhaps this can be usefully predicted from prior fully automatic experiments with the same topics, but that question would require further exploration before it could be used as a basis for study design.

3.7. Subjective reactions

Our analysis of participants' subjective views is based on the questionnaires and interview responses. On average, participants rated user-assisted condition as more easily used to find relevant documents than the automatic condition (4.0 vs. 3.5 on a 1–5 Likert scale, where 5 indicates easiest). We should note, however, that our study participants were not told which documents CLEF assessors judged to be relevant because the assessors could not complete their work until the participants selections from the user studies became available. This self-report data on retrieval effectiveness was therefore grounded only in the perception of the participants during the study. In response to more specific questions, all participants reported positively about the usefulness of user-assisted query translation. When asked to choose between a system with a user-assisted query translation function and a system without, most participants (14 of 20) preferred the user-assisted system. The two most commonly mentioned reasons for that preference were that all participants (20 of 20) felt that they could (somewhat or very) confidently select and/or deselect translations, and that it was (somewhat or very) useful to have the ability to modify the system's translation selections. In some sense, they preferred the user-assisted condition because they had more control over the system. Regarding usability more generally, participants reported that they found the user-assisted query translation system and the fully automatic system equally easy to use. Moreover, they perceived an equal need to reformulate their initial queries with both systems.

We must be cautious when interpreting self-report data because is was clear to the participants that the researchers they were working with had created the user-assisted query translation system in the hope that it would be useful. In such cases, there can be a natural tendency to focus on its advantages. Their judgment is, however, consistent with evidence from objective measures. Over 16 topics (counting Study 1 and Study 2 topics separately because of the differences in the document collections), the user-assisted condition resulted in an improvement in retrieval effectiveness over the automatic condition in 8 cases, a reduction in retrieval effectiveness in 5 cases, and no difference in 3 cases (with strict judgments; for loose judgments, the corresponding numbers are 8, 4, and 4). When the objective and subjective evidence is considered together, we conclude that a well designed facility for user-assisted query translation can sometimes be a useful capability in an interactive CLIR system.

3.8. Searcher behavior

A ‘‘search strategy’’ refers to a plan that a user constructs to guide their search process (Bates, 1979). 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 users of search engines seem to naturally develop without formal training (Marchionini, 1995). 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 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 blocks’’ 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 (three library science students and three others, out of a total of eight participants in that study) employed a building blocks variant. The building blocks strategy taught to librarians results in construction of a single Boolean query in conjunctive normal form (an AND operator across facets, with nested OR clauses to match the expected 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.

3.8.1. Source of initial 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 for topic 7. In another example, one participant used ‘‘*CGI*’’ and another used ‘‘*pixar*’’ in their initial query for a topic 8 (about computer animation in films) from their background knowledge. A third source of terms for the initial query was linguistic knowledge of synonymy, abbreviations and morphological variants. For example, ‘‘*anti-racism anti-prejudice*’’ appeared in a query for a topic in which the topic statement contained ‘‘*against racism*’’.

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 therefore focus on each condition in turn.

3.8.2. Searcher behavior with automatic translation

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 searcher 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 topic 8, the searcher 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.

3.8.3. 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 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. 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 search for topic 8, one participant first entered the query "*movie film computer animation CGI.*" They then removed several unwanted translations, but before clicking the "`Search`" button they 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 in Study 3 simply chose one displayed Spanish translation for each word ("*europeo*" for European and "*sindicato*" for union) and typed those translations 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. 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, (e.g., correcting "*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.

3.8.4. Factors affecting translation selection

As mentioned in Section 3.6, 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-min search sessions, the average was 40%, whereas in the third study, with 10-min search sessions, the average was only 18%.

Topic difficulty (indicated by a relative paucity of relevant documents in the collection) also seemed to affect the use of user-assisted query translation. As explained in Section 3.5, the collection being searched in study 2 was a strict subset of the collection searched 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 17% to 19%. We interpret this as an additional source of support for a claim 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.

Two trends are evident from these observations: (1) user-assisted query translation was used repeatedly, and (2) the participants in our study found ways of using it that we had not anticipated. Taken together, these suggest that further investments in refining the implementation would be worthwhile. That is the focus of the next section.

4. Building a better MIRACLE

The results from these user studies suggest that the participants generally appreciated the availability of user-assisted query translation, that they used the capability extensively, and that its use often yielded improved retrieval effectiveness when substantial numbers of relevant documents were available to be found and enough time was available to find them. Adverse effects on retrieval effectiveness were also noted, but they were less common than beneficial effects, and they were disproportionately associated with time-constrained or quantity-constrained searches. Additional studies will be needed before we can 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. But we are now able to see several productive directions for further work, and our results do seem to indicate that further work is warranted. In this section, we describe our progress since completion of these studies and we identify what we see as the most pressing open questions that merit further study.

4.1. Progressive refinement

Performing even the simplest search with our initial MIRACLE system required that the searcher first select "`Translate`" and then select "`Search`". Query translation was typically quite fast, and we observed that searchers would sometimes select "`Search`" without first examining the translation selection area. This way, they could examine some initial results before trying to make sense of the available translations for each query term. Because we used Pirkola's structured query method, which has been shown to make good use of all alternatives in manually created bilingual dictionaries, this sometimes worked well enough that translation deselection was not needed. As described in Section 3.6, we had also noted that searchers performed translation deselection or reselection more often in the second half of the longer sessions. This suggested that fluidly moving between automatic and user-assisted processing might be beneficial. We therefore redesigned the interaction strategy for our improved version of MIRACLE to search immediately and then allow translation deselection and reselection to progressively refine the search results, and we changed the implementation so that searchers can easily hide the translation selection window to gain more screen space for the automatic search results (and easily restore it later in their search process).

Implementation of this new capability proved to be fairly straightforward. When the client provides a query, the server formulates a structured query using InQuery syntax and stores that query in a file. The server then initiates an InQuery search, which reads the query from the file and writes a TREC-format ranked list. While the search is being performed, the server also informs the client of the known translations for each query term; these therefore typically become available for display and selection before the initial search completes. When the search completes, the server parses the results file, creates query-focused summaries for the top 10 search results, and then passes those summaries to the client for display. If the searcher elects to deselect some translation and then chooses "`Search Again`", the client informs the server which translations remain selected and the client then forms a new query and repeats the process.

In our original MIRACLE implementation, we had treated each query *de novo*, initially selecting all possible translations for each query term. We learned during the interviews that we conducted that searchers found this behavior to be counterintuitive and that they believed that retaining translation selections across queries would reduce their workload. This makes sense from the perspective of the widely used heuristic that homonymous terms typically exhibit only a single sense per discourse. If we consider a sequence of queries on the same topic as a single discourse, then deselecting translations for an inappropriate sense of some query term can quite plausibly be treated as a persistent action. We therefore retain translation selections across queries in our improved version of MIRACLE. Of course, this introduces the possibility that a searcher might deselect some translation when working on one topic that they later wish to select when working on another. Further user studies will be needed to determine whether persistent selection results in disorientation for users working over extended periods. If it does, we might also want to add a "`Reset`" function to allow the searcher to easily place the system back in a known initial state.

This new design has the advantage of simultaneously providing searchers with feedback on the quality of their query from two sources: the initial search results, and the available translations. A second benefit of the design is that the automatic condition simply becomes a special case of the user-assisted condition. We take advantage of that by providing a function that allows the user to hide or display or hide the translation selection area with a single mouse click. Hiding the translation selection area yields a simple interface that will appear familiar to novice searchers who are familiar only with Web search engines in which most of the screen real estate is devoted to summaries of highly ranked documents. The ability to toggle between the two conditions is also potentially useful to expert searchers, allowing them to easily reallocate screen space between user-assisted query translation and examination of document summaries.

4.2. Leveraging statistical translation lexicons

Pirkola's structured query method reduces the contribution of query terms that have at least one commonly-occurring translation in the collection being searched (Levow et al., 2005). This works well when manually prepared dictionaries are used. The most remarkable advance in CLIR research over the period spanned by our user studies was the clear emergence of statistical techniques based on large collections of translation-equivalent parallel text as the preferred source for translation mappings. All manually created dictionaries are incomplete (e.g., missing some less common translations and some important domain-specific terms and proper names). Automated CLIR systems using manually prepared dictionaries therefore typically yield 70–80% of monolingual mean average precision. Statistical translation lexicons now routinely achieve better than 90% of monolingual mean average precision when trained on large domain-appropriate collections of parallel text, and techniques for affordably generating such collections have been demonstrated (Yarowsky, 2003). We therefore placed a high premium on incorporating statistical translation lexicons into MIRACLE.

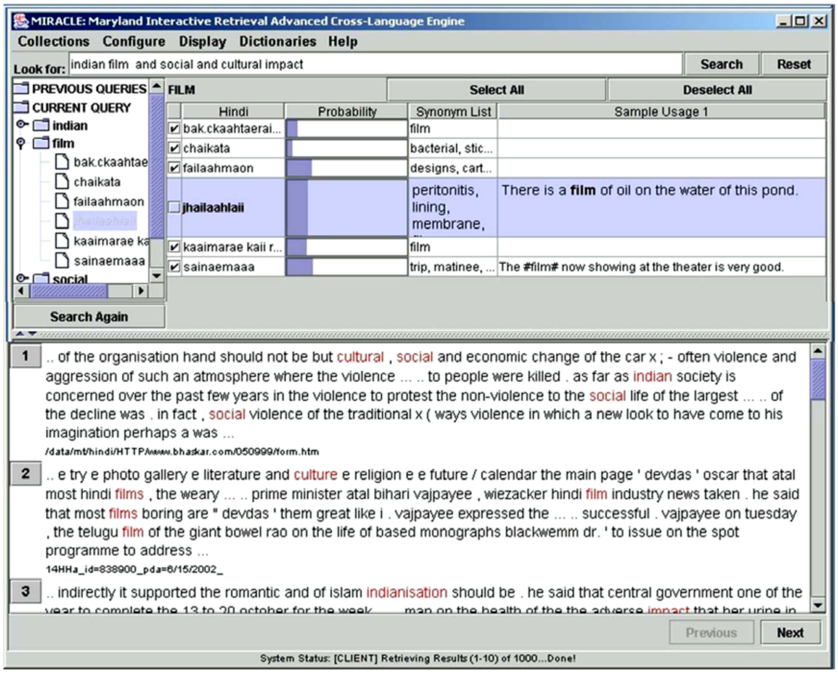
Statistical techniques for learning translation mappings introduce additional challenges, however, since in that case we learn not the presence or absence of a translation relationship, but rather the empirical probability of seeing those words together (in plausible locations) in a pair of sentences produced by professional translators. This typically yields translation mappings that couple a few highly probable translations with a large number of possibilities that occur with very low probability. Applying Pirkola's method across the full set of possible translations in such circumstances can yield quite poor results, since very large sets of possible translations are likely to include at least one term that is very common. One way of addressing this problem is to scale the TF and DF contributions of each candidate translation by the calculated probability of that mapping. The Perl Search Engine (PSE) implements that "weighted structured query" approach (Darwish & Oard, 2003; Wang & Oard, 2006). Incorporating PSE into MIRACLE required only minor query reformatting (to include probabilities with each term); otherwise, the file-passing process is identical to that used with InQuery.

4.3. Simultaneous display of multiple cues

One early design goal for MIRACLE was to present users with alternative cues for the meaning of possible translations of query terms, but in our original design they could only see one type of cue at a time. For example, back-translations were available by selecting one tab, while examples of usage were shown when a different tab was selected. It turned out, however, that different cues could often provide complementary evidence, and we noted that participants in our user study would sometimes rapidly flip back and forth between tabs in order to get the full picture. The most radical change that we made to the MIRACLE interface was to permit the simultaneous display of multiple cues. This naturally led to the table layout shown in Fig. 10. That allows the searcher to see at a glance which types of cues are available. As illustrated by the fourth row in that figure, moving the mouse over any row in the table temporarily enlarges that row enough to display full (word wrapped) entries.

An additional source for examples of usage, dictionary entries, was also added to MIRACLE as a result of our experience with Hindi (He et al., 2003). None of our available online dictionaries for German or Spanish had provided examples of usage in English for the non-English terms, but one such dictionary was available for Hindi. Its use proved to be quite straightforward. Moreover, no single source for examples of usage can reliably cover every plausible translation mapping that might be obtained from multiple sources, so the present version of MIRACLE displays examples of usage in multiple source-specific columns. An alternative design might prioritize those sources and then display the best example(s) (e.g., from dictionaries when possible, parallel text otherwise, and comparable text only when necessary).

The availability of translation probabilities made it possible to add another new type of cue to help with translation selection: an iconic representation of how commonly each translation is used. We show this as a horizontal bar with a length proportional to the translation probability, and we suppress the display of any translation with a probability less than some small threshold (0.01 in our present implementation). When several translation alternatives are shown, this approach allows the searcher to focus their attention on those that will have the greatest impact on search results.



Look for: indian film and social and cultural impact

	Hindi	Probability	Synonym List	Sample Usage 1
☑	bak.ckaahtaerai...		film	
☑	chaikata		bacterial, stic...	
☑	failaahmaon		designs, cart...	
☐	**jhailaahlaii**		peritonitis, lining, membrane,	There is a **film** of oil on the water of this pond.
☑	kaaimarae kaii r...		film	
☑	sainaemaaa		trip, matinee, ...	The #film# now showing at the theater is very good.

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.bhaskar.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=838900_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!

Fig. 10. The improved MIRACLE interface, configured for Hindi.

4.4. Search history

Displaying the translated query in the first version of MIRACLE consumed a significant amount of screen real estate. Over the course of our user studies we saw little evidence that users found this information to be helpful. We did, however, observe that searchers would often iteratively improve a query until they had adequately mined one part of the search space and then return to one of their previous queries and iteratively evolve from there towards some different goal. We have previously observed similar patterns in (monolingual) Web search engine query logs, and search history tools have been shown to be useful in some applications (Komlodi, Soergel, & Marchionini, 2006). We therefore redesigned the translated query display as part of a more compact and capable search history that now appears to the left of the translation selection area.

The key to this redesign was to view the search history as a three-level hierarchy, with the history being an ordered set of queries, each query being a set of terms, and each term having an associated set of translations. A tree browser (similar to the type used in Microsoft's ''Windows Explorer'') offers a familiar and flexible interaction metaphor for such a hierarchy. When users drill down to see translations of individual query terms, those that are (persistently) deselected are indicated with a gray box, as illustrated by the fourth translation of ''film'' in Fig. 10.

One useful side effect of this design is that the vertical orientation of expanded terms corresponds with the vertical stacking of those same terms in the translation selection area. In future designs we may want to couple those two elements even more closely, extending the entries in the tree to include the full content of the translation selection area. We did not do that for the present version of MIRACLE simply because the existing tree and table layouts in Java can adequately support our present goals with less implementation effort.

4.5. Other enhancements

We also made a number of more minor improvements to capture other lessons that we had learned in our user studies. One simple change was repetition of the query term being translated among the list of back-translations. The original query term is, of course, always a back-translation of itself, but we had omitted it in our original design because its presence would provide no new information to the searcher. We learned during the training that we provided to each participant that they wanted to understand the source of that information. When we explained how it was obtained, they often asked why, if our explanation was correct, the original term did not appear in the list of back-translations. We have found that retaining the original term in that list helps people to understand what they are actually looking at, and our improved MIRACLE system implements back-translation in that way. When designing IR systems, we naturally tend to focus on effectiveness. This incident points up the importance of focusing on the design of explainable tools as well, since searchers can better control tools that they understand.

In 2003, the Defense Advanced Research Projects Agency (DARPA) conduced two "surprise language" exercises in which research teams were challenged to apply language technologies to previously unforeseen languages in a brief period (initially, 10 days for Cebuano; later, 29 days for Hindi). Extending MIRACLE to Cebuano proved to be quite straightforward, but incorporating Hindi proved to be a greater challenge. A proliferation of character sets, with different sources of Hindi text typically using different proprietary encodings, meant that a translation lexicon learned using text from one source could not easily be used with text from a different source. Development of conversion tools for a substantial number of high-volume sources resolved that problem within a couple of weeks (Khudanpur, 2003). One serendipitous byproduct of that effort was that the standardized form used ASCII characters in a way that (approximately) preserved the phonetic representation of the original Hindi word. We displayed this transliterated form in MIRACLE as the candidate translation rather than showing the original Hindi, thus allowing the searcher to sound out the term in order to recognize loan words and proper names. The last translation shown in Fig. 10 illustrates one case in which this proved useful ("*sainaemaaa*" for the "*cinema*" sense of "*film*").

Finally, we also added a compact status display at the bottom of the screen to help searchers distinguish between occasional delays due to network latency and more serious problems such as a server crash. This also provides a convenient way of helping searchers to develop a richer mental model of system operation. For example, our present system always finds the top 1000 results and begins by displaying the first 10; status messages with that information help searchers to recognize that those numbers do not vary in response to their query.

4.6. Initial experience

Our improved version of MIRACLE has been used in two settings to date. The first was a formative evaluation conducted in 2003. Timed 10-min searches were performed by the first author of this paper (a native speaker of English with no ability to read Hindi) for 15 English topics in a collection of 41,697 Hindi documents, both of which were provided by the US National Institute of Standards and Technology (NIST). Details on the system configuration for this task are available in He et al. (2003). Relevance judgments made by our one searcher using MIRACLE were compared with judgments made by Hindi speakers using a pooled assessment methodology by the Linguistic Data Consortium. The resulting $F_{\beta=2} = 0.45$, which compares favorably with the best results obtained under similar time pressure in Study 3 ($F_{\beta=2} = 0.31$ for the user-assisted condition with loose relevance judgments), although differences in both the test collection and the searcher's understanding of the system design make it difficult to read much into such comparisons. We can, however, say with confidence that the improved version of MIRACLE can be used to find relevant documents in a language that the searcher cannot read. For example, in this study the searcher found a total of 151 documents that they believed to be relevant, 110 of which were assessed as relevant by a native speaker, equating to a precision (averaged over 15 topics) of 0.68.

In 2004, the improved MIRACLE system was used for a cross-language question answering task in an iCLEF user study. Eight native speakers of English each sought the answer to 16 questions in a Spanish document collection. The single best answer in each case (in the opinion of the searcher) was recorded on paper in English, hand-translated to Spanish by the native speaker who would perform the assessment, and then judged (by CLEF assessors) using procedures similar to those used for evaluation of automatic question answering systems. The recorded answer was judged to be correct in 79 of 128 cases (62%), providing further evidence that our improved version of MIRACLE can be used for practical tasks. Details on the system configuration for this task and preliminary analysis of the results can be found in (He, Wang, Luo, & Oard, 2004).

4.7. Next steps: better support for query formulation

Our focus in designing MIRACLE was to help searchers understand, and therefore mitigate, the effects of translation ambiguity. An obvious extension to this strategy would be to help users express their queries in ways that minimize the occurrence of ambiguity in the first place. It is well known that phrases exhibit far less translation ambiguity than individual words, and CLIR experiments in non-interactive settings have shown that phrase translation can yield improved retrieval effectiveness (Ballesteros & Croft, 1998). User studies by Lopez-Ostenero and his colleagues have found that phrases alone can adequately support some document recognition and query reformulation tasks (Lopez-Ostenero et al., 2002). Together, these results suggest that integrating phrase translation into MIRACLE might be useful.

MIRACLE could easily allow the searcher to manually mark phrases that they wish to have translated as a unit (e.g., by using quotation marks). Bilingual dictionaries normally show only translations for individual words, but the translated terms are sometimes multi-word expressions. Reversing the translation mapping is therefore one possible source for phrases that are candidates for translation. The recent introduction of "alignment templates" in statistical machine translation systems (Och & Ney, 2004) provides an alternative source for translations of phrases. Lopez-Ostenero introduced a third technique, first generating every possible translation of each constituent word and then filtering the result (in an order-independent manner) using co-occurrence counts in a large collection of text in the target language.

The computational aspects of phrase translation are now fairly well understood, but we do not yet have much experience with these capabilities in interactive settings. Lopez-Ostenero used phrases alone, while in our work we have translated only single words. It remains to be seen what searchers will do when they have these capabilities available together. In monolingual settings, searchers typically understand that placing quotation marks around a sequence of words requires that they appear together and in order in the documents that are being searched. But that is not true in CLIR; multi-word expressions can translate to single words, and multi-word expressions in the target language may be in an order different from the order of those words in the query. Moreover, the translation system used to display documents (and summaries of those documents) to the searcher may not translate the document-language terms that match query phrases back into those same phrases. User studies will be needed to explore these issues, so incorporating phrase translation into MIRACLE deserves high priority.

Our support for query reformulation is also in need of enhancement. Searchers often acquire vocabulary that can be used to formulate better queries by examining documents (or summaries of those documents) that they find in their initial searches. In monolingual applications, it then suffices to type that new term into a new query. Indeed, that simple strategy works so well that monolingual searchers rarely employ "relevance feedback" techniques that seek to automate the process (Spink & Jansen, 2004). The situation in CLIR applications is somewhat more complex, however. Two problems arise. First, retyping the term results in loss of information. Some specific term in the document gave rise to the English term that the searcher saw. When the searcher retypes that term, however, translation ambiguity would be unnecessarily reintroduced. Second, asymmetries in the translation resources used by the query and document translation components might result in the document-language term that gave rise to what the searcher saw not being generated as a possible translation of the query term that they retyped.

From the perspective of the system, these problems are entirely unnecessary; query enhancement using relevance feedback can be performed using only document-language terms. We should therefore extend MIRACLE to permit users to drag terms into their query from documents or from summaries. The English term that is copied will in essence simply be a label for the document-language term that generated it; we can indicate that to the user by initially deselecting other possible translations of that query term. This design raises three implementation challenges: (1) translation mappings must be available from the document translation process as a term-by-term mapping, (2) interface affordances for copying multi-word expressions may be needed, and (3) replication of a query term might now yield different translation selections. The first of these will limit our implementation flexibility considerably; although machine translation systems necessarily generate such mappings internally, off-the-shelf systems rarely expose such details to downstream applications. Fortunately, state-of-the-art statistical machine translation systems that can affordably be built by even fairly small research teams now rival the accuracy of existing commercial systems, so it should be practical to obtain the needed mappings using machine translation systems that are custom-built with this requirement in mind.

4.8. Next steps: better support for examination and use

Integrating state-of-the-art statistical machine translation systems could pay dividends in other ways as well. For example, translation mappings could facilitate more precise highlighting of document terms that match the searcher's query. Perhaps most importantly, the introduction of affordable automated measures for translation accuracy has resulted in rapid progress in statistical machine translation over the past few years. There is not yet any indication that rate of progress is abating, so the investment needed to more closely integrate statistical machine translation into interactive CLIR systems could pay increasing dividends as improvements in translation quality begin to approach the point where concept learning can be adequately supported from foreign-language texts. We have recently started some initial experiments to assess the ability of searchers to make sense of statistical translations (He et al., 2006), and our preliminary results seem encouraging. The ability to affordably create translation systems for new languages also holds promise for extending the capability for interactive CLIR beyond the relatively few "wealthy" languages to which it has been applied to date. As our "surprise language" experience has shown, statistical machine translation systems can be adapted to new languages relatively easily (and thus, relatively inexpensively), a useful characteristic for a world in which thousands of languages are in common use.

Another possible investment with potential for substantial impact would be integration of improved summarization techniques. Our present keyword-in-context technique helps the searcher understand why a document was ranked highly by showing brief passages that match with several query terms, but at the cost of displaying a substantial amount of text. Researchers in computational linguistics have begun to focus on extending automated summarization techniques to cross-language applications (Zajic, Dorr, Lin, & Schwartz, 2005), so this would be a propitious time to re-examine that aspect of the MIRACLE system design. For example, translation quality might be improved by leveraging measures of translation confidence when selecting passages to display, and brevity might be improved using linguistically-guided trimming. The acid test for both machine translation and automatic summarization are the extent to which they can support real people performing real tasks. Coupling the best of those technologies with state-of-the-art techniques for interactive CLIR thus creates a symbiosis in which the CLIR system provides an evaluation venue that can help stimulate further advances in translation and summarization technology, with those advances in turn helping us to build even better systems for interactive CLIR.

5. Conclusion

The culture of IR evaluation values objective measurement for good reason; many ideas that initially appear promising do not survive their first encounter with an evaluation framework that models a real task. The design of our user studies reflects that influence; we structured our studies around a quantitative comparison of relative retrieval effectiveness with and without user-assisted query translation. Because we adopted a mixed-methods design that drew on both quantitative and qualitative methods, we uncovered a broader range of issues than would have been possible with quantitative methods alone. Ultimately, this provided the basis for the design of an improved MIRACLE system that others can now use in their own research.

Our results point to several important conclusions, including:

- Searchers are able to find topically relevant documents in a language that they cannot read. Specifically, they find many of the same documents that a searcher skilled in the document language would find when performing a monolingual search. Moreover, they are able to perform at least one actual task (answering factual questions) correctly using such a system in more than 60% of the cases.
- User-assisted query translation provides a capability that was used repeatedly and remarked upon positively by study participants. Use of this capability was seen to be more often helpful than harmful in this study, although the lack of statistical significance indicates that outcome should be taken as suggestive rather than conclusive.

- Introduction of user-assisted query translation has implications for the iterative search process that interactive searchers employ. These process implications, in turn, have implications for the design of interactive search systems. Examples include progressive refinement (grounded in an observed preference for viewing search results before refining translation choices) and search histories (grounded in an observed tendency to use a previous query as an anchor for a new episode of iterative refinement).
- Support for interaction can be extended fairly easily to accommodate new languages. One of the hallmarks of research on automated techniques for CLIR has been the relative ease with which new languages can be introduced. We have shown that coupling our corpus-based techniques for constructing examples of usage with back-translation and statistical machine translation can yield credible interactive search systems at modest effort and expense.
- Our qualitative analysis of the studies show that studying the process(es) by which CLIR machines are used is as important as examining the effectiveness of those machines in producing desired results. Many useful insights about the employment of user-assisted query translation resulted from examining the actual behaviors of our users. Our results show that searcher behavior changed even over the short time span of a single half-day session. Participants learned from the interactions, adapted to the capabilities of the machines, and developed new search strategies. Therefore, the design of CLIR machines should aim to help people develop effective strategies, and the evaluation designs should take this adaptation into consideration.

Of course, much still remains to be done. For example, we have focused on how searchers learn to refine their queries, paying little attention to the equally important question of how they will refine their own understanding of what they are really looking for. Our reason for that was simple: recognized deficiencies in present machine translation systems at the time we conducted our studies made reading complex foreign-language documents a frustrating and time consuming task. As machine translation capabilities improve, longitudinal studies of searchers working with interactive CLIR systems over extended periods will become increasingly important. But we need not wait for improved translation to do that; we could today design studies in which searchers consult monolingual sources to extend their understanding, switching to CLIR once they have a good sense for what they are looking for. Exploring questions that turn on the evolution of internalized information needs would require a study design quite different from what we employed for our more narrowly focused questions. Extending MIRACLE to accommodate same-language searching as well as CLIR would offer a way to begin exploring these questions.

Cross-language information retrieval has sometimes been referred to as “the problem of finding documents that you cannot read,” with the implication that doing so might be of debatable value. The same formulation for within-language search, however, would be “the problem of finding documents that someone happened to write in the same language as your query.” The debate need not turn solely on whether you can read what you find. Rather, the question to be answered is whether you can afford not to even know what exists in other languages. Perhaps that question could have been answered affirmatively in the past, but it seems unlikely that the 21st century will be as tolerant of such myopia.

Acknowledgements

The authors would like to thank Julio Gonzalo and Fernando López-Ostenero for helping to coordinate the CLEF interactive track, for which the first author of this paper was a co-organizer, and Jun Luo, one of the key MIRACLE makers. This work has been supported in part by DARPA cooperative agreement N660010028910.

References

Adriani, M. (2000). Cross-language retrieval experiments at CLEF-2000. In *Proceedings of evaluation of cross-language information retrieval systems: First workshop of the cross-language evaluation forum.*

Ballesteros, L., & Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th international ACM SIGIR conference on research and development in information retrieval.*

Ballesteros, L., & Croft, W. B. (1998). Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 64–71). ACM Press.

Bates, M. J. (1979). Information search tactics. *Journal of the American Society for Information Science, 30*(4), 205–214.

Braschler, M. (2004). Combination approaches for multilingual text retrieval. *Information Retrieval, 7*(1–2), 183–204.

Capstick, J., Diagne, A. K., Erbach, G., Uszkoreit, H., Leisenberg, A., & Leisenberg, M. (2003). A system for supporting cross-lingual information retrieval. *Information Processing and Management*, 275–289.

Darwish, K., & Oard, D. W. (2003). Probabilistic structured query methods. In *Proceedings of the 21st annual 26th international ACM SIGIR conference on research and development in information retrieval* (pp. 338–344). ACM Press.

Demner-Fushman, D., & Oard, D. W. (2003). The effect of bilingual term list size on dictionary-based cross-language information retrieval. In *36th annual Hawaii international conference on system sciences, Hawaii.*

Dorr, B. J., He, D., Luo, J., Oard, D. W., Schwartz, R., Wang, J., et al. (2003). iCLEF 2003 at Maryland: Translation selection and document selection. In *Fourth workshop of the cross-language evaluation forum, LNCS, Vol. 3236* (pp. 435–449).

Gao, J., Xun, E., Zhou, M., Huang, C., Nie, J.-Y., & Zhang, J. (2001). Improving query translation for cross-language information retrieval using statistical models. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 96–104). ACM Press.

Habash, N. Y. (2003). *Generation-heavy hybrid machine translation*. Unpublished doctoral dissertation, Department of Computer Science, University of Maryland, College Park.

He, D., Oard, D. W., & Plettenberg, L. (2006). Studying the use of interactive multilingual information retrieval. In *SIGIR workshop on new directions in multilingual information access.*

He, D., Oard, D. W., Wang, J., Luo, J., Demner-Fushman, D., Darwish, K., et al. (2003). Making MIRACLEs: Interactive translingual search for Cebuano and Hindi. *ACM Transactions on Asian Language Information Processing*, 219–244.

He, D., Wang, J., Luo, J., & Oard, D. W. (2004). iCLEF 2004 at Maryland: Summarization design for interactive cross-language question answering. In *Fifth workshop of the cross-language evaluation forum, LNCS, Vol. 3491* (pp. 348–362).

He, D., Wang, J., Oard, D. W., & Nossal, M., (2002). Comparing user-assisted and automatic query translation. In *Third workshop of the cross-language evaluation forum, LNCS, Vol. 2785* (pp. 400–415).

Hull, D. A., & Grefenstette, G. (1996). Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval.*

Kang, B.-J., & Choi, K.-S. (2000). Two approaches for the resolution of word mismatch problem caused by English words and foreign words in Korean information retrieval. In *Proceedings of the fifth international workshop on information retrieval with Asian languages* (pp. 133–140).

Khudanpur, S. (2003, October). Hindi 101. Team TIDES Newsletter, 2–3. (http://language.cnri.reston.va.us/TeamTIDES.html).

Kim, T., Sim, C.-M., Yuh, S., Jung, H., Kim, Y.-K., Choi, S.-K., et al. (1999). From To-CLIR: Web-based natural language interface for cross-language information retrieval. *Information Processing and Management, 35*(4), 559–586.

Komlodi, A., Soergel, D., & Marchionini, G. (2006). Search histories for user support in user interfaces. *Journal of the American Society for Information Science and Technology, 57*(6), 803–897.

Landauer, T. K., & Littman, M. L. (1990). Fully automatic cross-language document retrieval using latent semantic indexing. In *Proceedings of the sixth annual conference of the UW centre for the new Oxford English Dictionary and text research* (pp. 31–38). Waterloo Ontario: UW Centre for the New OED and Text Research.

Levow, G.-A., Oard, D. W., & Resnik, P. (2005). Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management, 41*(3), 523–547.

Lopez-Ostenero, F., Gonzalo, J., Penas, A., & Verdejo, F. (2002). Interactive cross-language searching: phrases are better than terms for query formulation and refinement. In *Proceedings of the third cross-language evaluation forum.*

Marchionini, G. (1995). *Information seeking in electronic environments*. Cambridge University Press.

Maxwell, S. E., Dalaney, H. D., & Dimmick, J. W. (2003). Designing experiments and analyzing data: A model comparison perspective, 2nd ed. Lawrence Brhaum Assoc.

McNamee, P., & Mayfield, J. (2002). Comparing cross-language query expansion techniques by degrading translation resources. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 159–166).

Monz, C., & Dorr, B. J. (2005). Iterative translation disambiguation for cross-language information retrieval. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 520–527).

Oard, D. W. (2002). When you come to a fork in the road, take it! In *Working notes of the SIGIR workshop on cross-language information retrieval: A research roadmap.*

Oard, D. W., & Diekema, A. (1998). Cross-language information retrieval. *Annual Review of Information Science and Technology, 33*, 223–256.

Oard, D. W., & Ertunc, F. (2002). Translation-based indexing for cross-language retrieval. In *Proceedings of advances in information retrieval, 24th BCS-IRSG European colloquium on IR research* (pp. 324–333).

Oard, D. W., Gonzalo, J., Sanderson, M., Lopez-Ostenero, F., & Wang, J. (2004). Interactive cross-language document selection. *Information Retrieval*, 205–228.

Och, F. J., & Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics, 30*(4), 417–449.

Ogden, W., Cowie, J., Davis, M., & Ludovik, S. N. (1999). Keizai: an interactive cross-language text retrieval system. In *Machine translation summit VII, workshop on machine retrieval.*

Peters, C., Clough, P., Gonzalo, J., Jones, G., Kluck, M., & Magnini, B. (Eds.). (2004). *Multilingual information access for text, speech and images 5th workshop of the cross-language evaluation forum.*

Petrelli, D., Levin, S., Beaulieu, M., & Sanderson, M. (2006). Which user interaction for cross-language information retrieval? Design issues and reflections. *Journal for American Society of Information Science and Technology, 57*(5), 709–722.

Pirkola, A. (1998). The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st annual international ACM information retrieval* (pp. 55–63). Melbourne, Australia: ACM.

Resnik, P., Oard, D., & Levow, G. (2001). Improving cross-language retrieval using backoff translation. In *Proceedings of the first international conference on human language technologies, San Diego, California*.

Simon, H. A., Dantzig, G. B., Hogarth, R., Piott, C. R., Raiffa, H., Schelling, T. C., et al. (1986). Decision making and problem solving. In *Research briefings 1986: Report of the research briefing panel on decision making and problem solving*. Washington, DC: National Academy Press.

Spink, A., & Jansen, B. J. (2004). *Web search: Public searching of the Web*. Kluwer Academic Publishers.

Voorhees, E. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management, 36*, 697–716.

Wang, J., & Oard, D. W. (2001). iCLEF 2001 at Maryland: Comparing Word-for-Word Gloss and MT. In C. Peters, M. Braschler, J. Gonzalo, & M. Kluck (Eds.), *Evaluation of cross-language information retrieval systems: Second workshop of the cross-language evaluation forum, clef 2001. Darmstadt, Germany* (pp. 336–354).

Wang, J., & Oard, D. W. (2006). Combining bidirectional translation and synonymy for cross-language information retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development on information retrieval*.

Yarowsky, D. (2003). Technical note: Scalable elicitation of training data for machine translation. *TeamTides Newsletter*.

Zajic, D., Dorr, B. J., Lin, J., & Schwartz, R. (2005). UMD/BBN at MSE2005. In *Proceedings of the MSE2005 track of the association for computational linguistics workshop on intrinsic and extrinsic evaluation measures for MT and/or summarization*.