Support for Interactive Document Selection in Cross-Language Information Retrieval

Douglas W. Oard and Philip Resnik
University of Maryland

Corresponding author:

Douglas W. Oard

Digital Library Research Group

College of Library and Information Services

4121G Hornbake (South Wing)

University of Maryland

College Park, MD 20742-4345

Phone: (301) 405-7590

Fax: (301)314-9145

Running title: Interactive Selection in Cross-Language IR

Abstract

As digital libraries grow to global scale, the provision of interactive access to content in many languages will become increasingly important. In systems that support query-based searching, the presence of multilingual content will affect both the search technology itself and the user interface components that support query formulation, document selection, and query refinement. This article describes the interactions among these components and presents a practical way of evaluating the adequacy of the selection interface. A categorization-based model for the user's selection process is presented, and an experimental methodology suitable for obtaining process-centered results in this context is developed. The methodology is applied to assess the adequacy of a selection interface in which multiple candidate translations for a term can be simultaneously presented. The results indicate that the modeled selection process is somewhat less effective when users are presented with multi-translation glosses from Japanese to English rather than materials generated originally in English, but that users with access to the gloss translations substantially outperform a naive Bayesian classification algorithm.

Introduction

Modern research libraries must acquire, catalog, store, locate, and provide information that is expressed in a wide variety of languages, scripts, and media in order to respond to the needs of researchers that use their facilities. As the increasingly interdependent global information infrastructure expands, interactive access to multilingual content will likely become an increasingly important feature. Perhaps even more important, as digital library technology for organizing and providing access to information becomes an integral part of this global network, the paradigm shift toward end-user searching that is currently evident in monolingual applications can be expected to extend to cross-language searching as well. Providing well designed tools to support that process will require integrating three disciplines: information retrieval, library science, and machine translation.

Query-based information retrieval systems pose a co-design problem among the three components shown in Figure 1: support for the user's query formulation and refinement processes, the search technology that accepts the query and returns a retrieved set of manageable size, and support for the process by which users select documents in the retrieved set that merit further examination. This perspective is grounded in an understanding of information seeking behavior that informs the design of conventional libraries today. Taylor described the process by which users compromise their initial *visceral* information need to conform to both their own present perception of that need (the *conscious* information need), their expressive abilities (which produce a *formalized* information need) and the perceived capabilities of the information provider that they will consult (for which they craft a *compromised information need or* query) (Taylor, 1962).

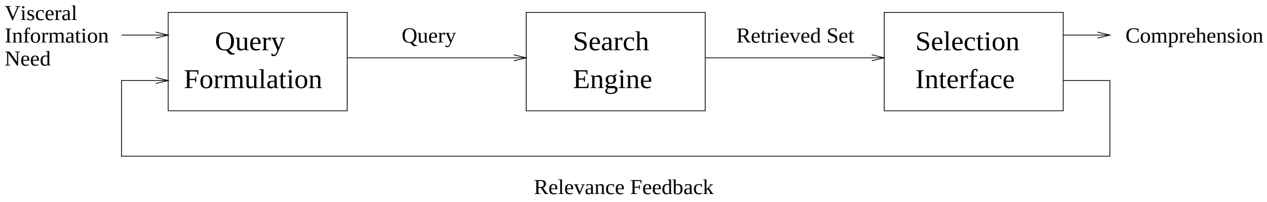


Figure 1: An interactive information retrieval process.

Search technologies typically use a combination of word matching and shallow natural language processing techniques such as morphological analysis to determine which documents in a collection should be retrieved, and the retrieved set is then presented to the user through some sort of selection interface. Typically document titles and other identifying information are ranked in an order approximating decreasing likelihood that each document will satisfy the information need that motivated the query. Often users are given the opportunity to disaggregate such lists, examining the full text of individual documents. Hearst and Pederson have shown that clustering documents in the retrieved set can also be useful (Hearst & Pederson, 1996), and clustering interfaces are beginning to appear in search systems for the World Wide Web in an effort to provide meaningful responses to short queries.

Regardless of how the results are displayed, the selection interface must support two types of decisions. The first is a selection decision, when the user must decide whether comprehension of a document's contents would serve the user's intended purpose. The second is a refinement decision, continuing an iterative process by which users seek to move their query closer to the visceral information need that motivates their search. This can be accomplished via manual query refinement, in which users examine the retrieved set and craft a more suitable query (perhaps simultaneously modifying their conscious information need); it can also be supported by the "relevance feedback" loop shown in Figure 1, in which the user designates good exemplar documents in the retrieved set and then the system automatically uses terms from those documents to augment the original query.

Because the components shown in Figure 1 are closely coupled, retrieval systems pose a co-design problem in which important capabilities can sometimes be provided more easily by one component than another. For example, although only a fraction of the top-ranked documents returned by a Web search system may actually be useful, a selection interface that helps the user rapidly recognize useful documents could lead to a system design that satisfies user needs. Similarly, users who are initially unable to form precise queries could initially pose a general query if the selection interface supported query refinement effectively. When designing information retrieval systems that support cross-language searching, it will be necessary to reexamine the way functions are divided among these three components. Consider, for example, a user who is unfamiliar with a language in which potentially useful documents are written. Clearly such a user could not be expected to reliably provide query terms that would appear in the same form in the documents — at least a portion of that role must be assumed by the automated system.

Because queries are typically much shorter than documents, for retrospective retrieval from large document collections it is generally more efficient to translate the query than to translate and store each document. And if queries must be accepted in several languages, the efficiency advantage of query translation over document translation increases. A number of automatic query translation techniques have been developed, and reported results suggest that between half and three quarters of the documents found using same-language queries can also be found using fully automatic query translation from another language (Hull & Oard, 1997). The most effective techniques combine information extracted from dictionaries with shallow natural language processing operations such as part-of-speech tagging and phrase recognition that serve to limit the effect of translation ambiguity. It seems reasonable to expect that designing the query formulation interface to support interactive disambiguation during query translation could produce even better retrieval effectiveness, and systems which support this approach are beginning to appear (Davis & Ogden, 1997).

Retrieval effectiveness will be of limited value if users are unable to interpret the contents of the retrieved document set, and this is an important limitation of the query translation approach that must be addressed in the selection interface. Before using individual documents to support query refinement or reading them for comprehension, experienced searchers typically apply sophisticated heuristics based on factors such as title, date, and authorship to identify promising documents. Systems that support cross-language searching based on query translation will need to incorporate machine translation in their selection interface to support these decision processes.

Although some users may ultimately require high quality translations of selected documents, it may be impractical to translate every document in the retrieved set sufficiently rapidly to support interactive browsing. For example, in our experiments for the Text REtrieval Conference (TREC-6) we have found that a typical newspaper story can be automatically translated from German to English in about a minute on a SPARC 20 (Oard & Hackett, 1997). The "TITAN" multilingual Web search system overcomes this limitation by translating only the titles on the Web pages in their returned set from English into Japanese using a translation technique that is tuned to the observed semantics of Web page titles (Kikui, Hayashi, & Suzaki, 1996). Since users of monolingual systems often examine the contents of some documents in the retrieved set when making query refinement and selection decisions, we believe that it is important to provide some means for browsing the documents themselves. In the next section we describe an interface for examining document titles or full documents that is based on rapidly constructed word-by-word gloss translations and present an experimental paradigm suitable for obtaining process-centered results in this context.

Process-Centered Evaluation of Gloss Translations

Gloss translations can potentially support both document selection and query refinement. The particular method we adopt is similar in spirit to the "Cliff-note" mode advocated by Church and Hovy, in which the original text is annotated with some word-by-word translations (Church & Hovy, 1993). Davis and Ogden have implemented a selection interface based on this principle for their QUILT cross-language text retrieval system, in which they replace words for which at least one translation is known with one possible translation and then allow users to select alternative translations using pop-up menus (Davis & Ogden, 1997). We take that approach one step further and simultaneously present up to three alternative translations when a single appropriate translation is not easily determined. We have observed that people can disambiguate words fairly easily in context, although this is a task that computers currently perform quite poorly. For example, the user presented with "Intel presente the new (flea, integrated chip) (from the, of the) Pentium Pro" has no difficulty ruling out the 'flea' interpretation of *puce*.[1] Because this "pop out" effect appears to be so pronounced, we believe that in this application it makes

[1] The source text for this example, in French, reads: *Intel presente la nouvelle puce du Pentium Pro.* (EDUPAGE, 2 November 1995, ⟨http://ijs.com/edupage/fr/⟩.)

sense to present alternatives where possible rather than to attempt disambiguation decisions automatically. When no translation is found in the dictionary, we present the untranslated word if the original language shares a common character set with the user's language. Otherwise we replace the word with ellipses ("...").

Arnold et al. suggest that evaluation of machine translation systems can be classified into *operational*, *declarative*, and *typological* criteria (Arnold, Sadler, & Humphreys, 1997). Operational evaluations take into account the end-user process within which the machine translation system will be used; the other two paradigms focus on subjective measures of translation adequacy (such as intelligibility and accuracy) and coverage of linguistic phenomena (via creation of standard test suites). A similar distinction is present in information retrieval evaluations, in which effectiveness measures such as recall and precision are used to characterize the effectiveness of the search technology component and user studies seek to reveal the interaction among all three components for the retrieval system. While declarative and typological evaluations of machine translation systems and recall-precision evaluations of information retrieval systems offer useful insights, there is no clear way to use the results of such evaluations to guide the design of components that depend upon user interaction. On the other hand, Arnold et al. point out that operational evaluations suffer a lack of generality, and they are costly and time consuming because they require putting the system "into the document processing environment in as realistic a manner as possible." Moreover, replicability by different researchers is a problem because two operational environments are rarely similar enough to afford a fair comparison between competing approaches to the same task.

As with declarative, typological and recall-precision evaluations, we have chosen to focus on a single component, in this case the selection interface shown in Figure 1, abstracting away from the full user environment. But unlike a recall-precision evaluation, in which strong assumptions are made about the selection process in order to obtain reproducible results, we have instead fixed the output of the search technology and sought to measure the effect of our gloss translation strategy on a specific decision process. This approach offers some of the advantages of operational evaluations while providing the opportunity to conduct the evaluation in a controlled laboratory setting. To the extent that the evaluation task is a fair abstraction of the true end-user process, the results obtained can inform design decisions for the overall system.

Our evaluation technique seeks to measure the effectiveness of gloss translations with respect to a decision making task. Design of the task was guided by the following considerations:

Minimizing a priori biases. In some experiment designs, each participant is provided with an identical statement of a formalized information need. There is, however, a large body of evidence that suggests that even expert searchers will differ on the interpretation of such a statement (c.f., (Wilbur, 1994)). Although it is obviously impossible for a set of experimental subjects to agree on a common visceral information need that motivates their search, experimental results would be more informative if the users were able to at least develop a common conscious information need. Doing so without using a formalized information need statement as a surrogate would also help to avoid incorporating the experimenters' perceptions or biases into the task.

Approximating real decisions. With a selection interface, users choose among actions on the basis of their conscious information need and their perception of the

> content of a retrieved item. The user may begin with a sense of what is "out there," and can refine that perception as they examine returned items. For example, a user searching for material on "language translation" might expect that material resulting from their search would pertain to commercial translation products, translator services, and technical papers on machine translation; looking at candidate results might lead to additional categories such as literature translations and Bible studies. Abstractly, at the point in time when an action is chosen for a given item, we model the user as possessing a set of conceptual categories, and the selection process can be viewed as "bucketing" the item into one of those categories and then choosing an action accordingly. This model leads naturally to a *categorization* paradigm, in which subjects in the laboratory setting perform a task analogous to that central first step.

The binary judgments of topical relevance used in recall-precision evaluations of search technology effectiveness are a simple version of this categorization paradigm in which the user has just two categories in mind, one characterizing sought-after information and the other representing "everything else," and the topic is explicitly stated as a formalized information need. We believe that for selection interface evaluations, a categorization model and our focus on the conscious information need better reflects the challenges that users face in their decision process.

Experiment

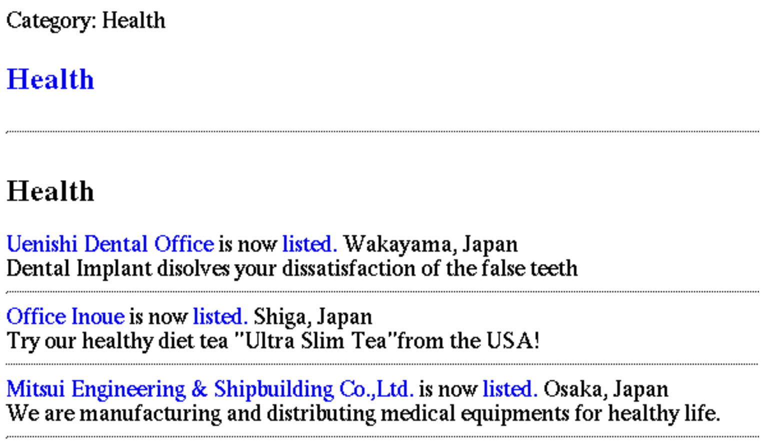
The experimental design is relatively straightforward; we define a task in which all subjects are faced with the same categorization problem. Some of those subjects are given materials in English to categorize that are drawn from a parallel collection in which every document appears in two languages. Other subjects are given the same *content* to categorize, but in the form of gloss translations into English from the other language. If the subjects given gloss-translated materials make decisions that are similar to those made by the subjects given the English materials (allowing for normal variability in people's judgments), we can conclude that the gloss translations are suitable for this purpose.

Materials. Experimental items were selected from the Nihongo Yellow Pages, a business directory site on the World Wide Web.[2] The site was chosen because it contains information across a variety of topic areas, because each business directory listing consists of a concise and informative description, and because most listings are available in both Japanese and English. In our experiments we used listings from the Nihongo Yellow Pages' Education, Finance, What's New, Entertainment, and Health categories, selecting a total of 73 business listings at random from those areas.

For each of these listings we created a 3 × 5-inch index card with a business advertisement in English and a corresponding card containing the same content as expressed in Japanese together with a gloss translation of that content. By way of illustration, Figure 2 shows three items in English, with their corresponding translated items appearing in Figure 3. Details of the translation process are given in (Resnik, 1997). Our current prototype handles gloss translation from Japanese, French, and Spanish to English, though our test collection for these experiments permits only Japanese-English evaluation.

Procedure. In order to create topical categories in an objective way, we randomly selected 32

[2] ⟨http://www.nyp.com/HTML/directory.html⟩



Category: Health

Health

Health

Uenishi Dental Office is now listed. Wakayama, Japan
Dental Implant disolves your dissatisfaction of the false teeth

Office Inoue is now listed. Shiga, Japan
Try our healthy diet tea "Ultra Slim Tea"from the USA!

Mitsui Engineering & Shipbuilding Co.,Ltd. is now listed. Osaka, Japan
We are manufacturing and distributing medical equipments for healthy life.

Figure 2: English items from Nihongo Yellow Pages

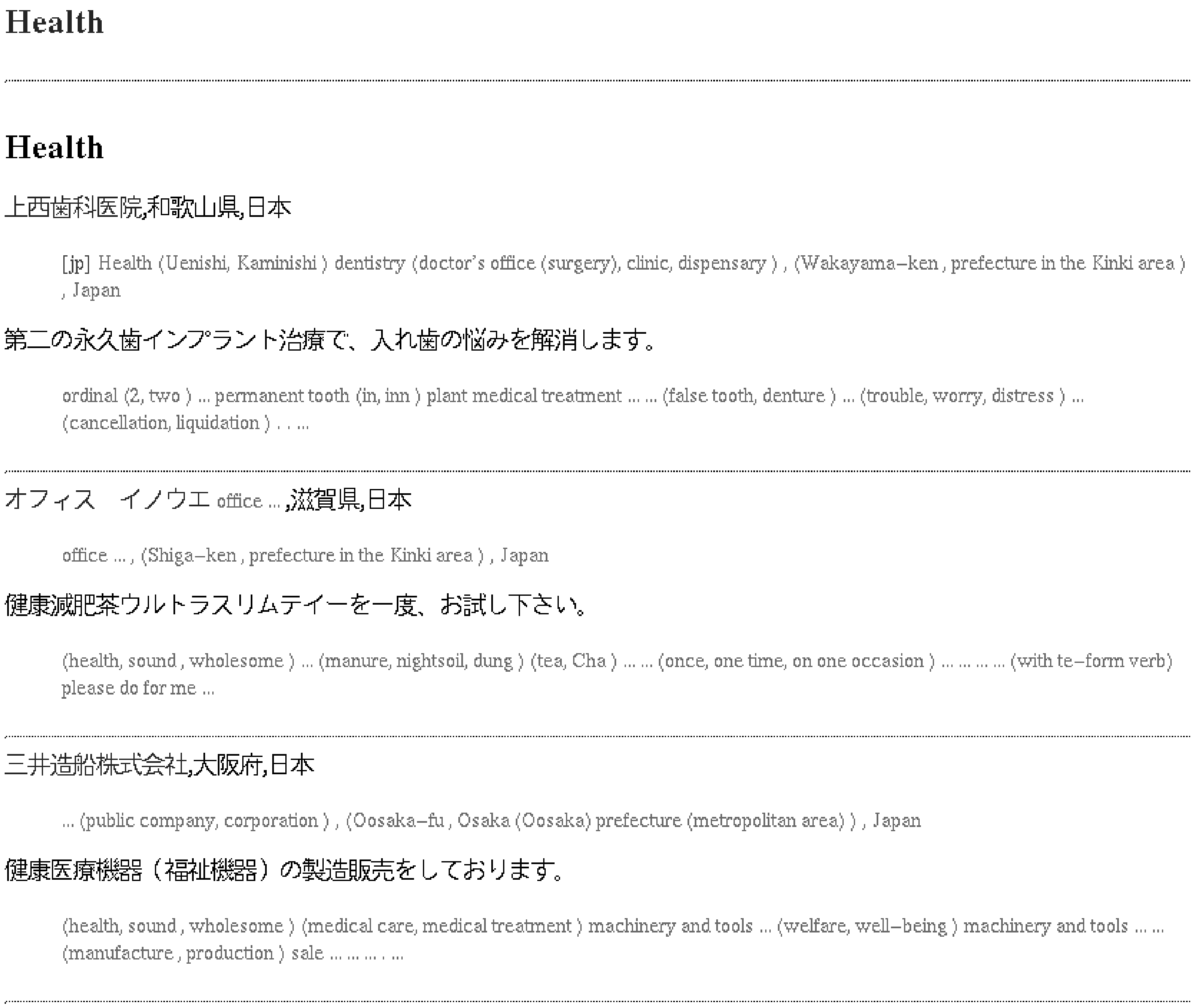
of the 73 English cards and gave them to 3 different subjects,[3] with instructions to sort the cards "into 4-6 piles of roughly equal size, placing cards in the same pile when you think they should 'go together', for example because they are related to similar topics." One subject created 4 piles, another 6, and the third 7 piles. We chose the 6 piles created by the second subject as defining the topical categories for the remainder of the study; the selection of categories is discussed further in the next section.

A set of 6 subjects participated in the control condition of the experiment, which involved categorizing material in English. The procedure had two parts (see Figure 4).

1. First, subjects were presented with the 6 piles of English cards created as described above. They were asked to read through each pile and decide "what you think each one is about." As a memory aid, subjects were encouraged to write a description of their choosing on a Post-It note for each pile, and place the note next to the corresponding pile.

2. Having formed their own impression of the 6 topical categories, subjects in the control

[3] All subjects in this first experiment were employees of Sun Microsystems in Chelmsford, Massachusetts, solicited as volunteers. All participants in both studies were fluent in English and nobody who saw gloss translations of Japanese materials was at all familiar with Japanese.



Health

Health

上西歯科医院,和歌山県,日本

[jp] Health (Uenishi, Kaminishi) dentistry (doctor's office (surgery), clinic, dispensary) , (Wakayama–ken , prefecture in the Kinki area) , Japan

第二の永久歯インプラント治療で、入れ歯の悩みを解消します。

ordinal (2, two) ... permanent tooth (in, inn) plant medical treatment ... ... (false tooth, denture) ... (trouble, worry, distress) ... (cancellation, liquidation) . . ...

オフィス　イノウエ office ... ,滋賀県,日本

office ... , (Shiga–ken , prefecture in the Kinki area) , Japan

健康減肥茶ウルトラスリムテイーを一度、お試し下さい。

(health, sound , wholesome) ... (manure, nightsoil, dung) (tea, Cha) ... ... (once, one time, on one occasion) ... ... (with te–form verb) please do for me ...

三井造船株式会社,大阪府,日本

... (public company, corporation) , (Oosaka–fu , Osaka (Oosaka) prefecture (metropolitan area)) , Japan

健康医療機器（福祉機器）の製造販売をしております。

(health, sound , wholesome) (medical care, medical treatment) machinery and tools ... (welfare, well–being) machinery and tools ... ... (manufacture , production) sale ... ...

Figure 3: Translated items from Nihongo Yellow Pages

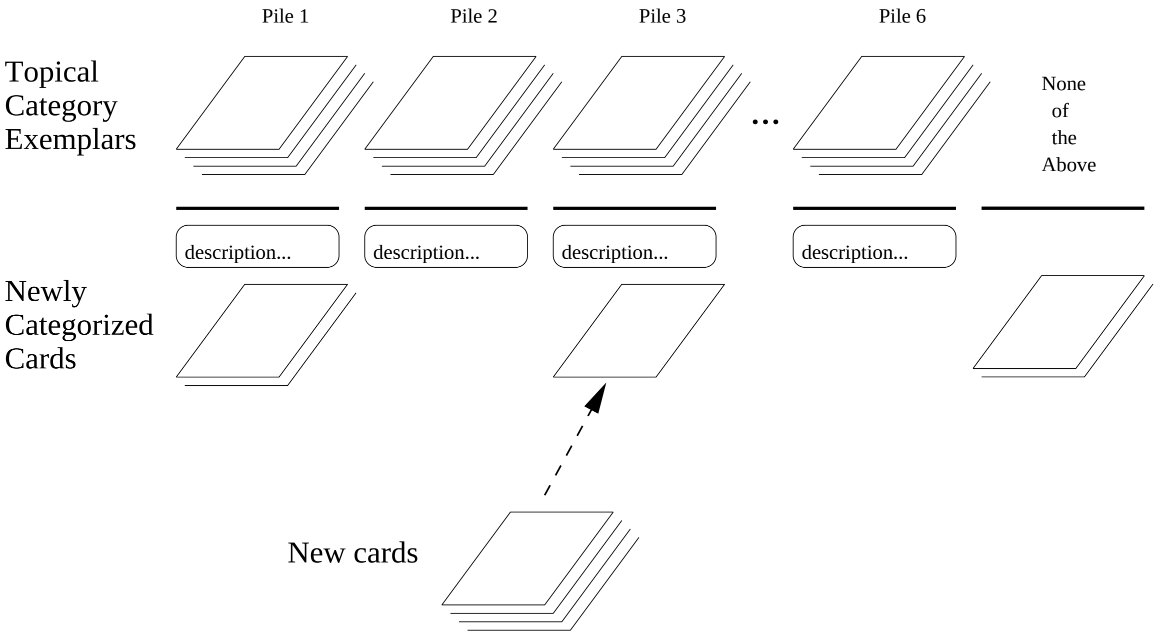


Figure 4: Categorization of new items

> condition were now given 32 new randomly-selected cards in English. They were instructed that for each new card, they should decide in which of the 6 categories it "belongs" and place it next to the corresponding pile. They were also given the option of placing cards in a seventh "none of the above" category.

Subjects were told to take as long as they liked on both parts of the categorization task.

A set of 8 subjects participated in the experimental condition. Part 1 of the experimental condition was completely identical to Part 1 of the control condition: subjects looked at exactly the same 6 piles of English cards and formed their own mental description of each topical category, writing down a short description as a memory aid. Part 2 was also identical, with one crucial exception: instead of being given cards in English to place into categories, subjects were given the corresponding cards containing gloss translations of Japanese, similar to the example shown in Figure 3.

In both control and experimental conditions, order of presentation for the cards to categorize was counterbalanced, with half of the subjects receiving cards in one random order and the other half receiving them in the reverse order.

Results. The categorization data gathered in the experiment were analyzed following the method of Hripcsak et al. (Hripcsak, Friedman, Alderson, DuMouchel, Johnson, & Clayton, 1998). In their study, they compared the performance of physicians, laypersons, and several computer programs on the task of classifying chest radiograph reports according to the presence or absence of 6 medical conditions. Our adaptation of their analysis is almost completely direct, with subjects in the control condition (English cards) corresponding to the physicians, subjects in the experimental condition (translated cards) corresponding to laypersons, and our automatic baseline runs corresponding to a subject in their baseline conditions, where they used simple keyword-based classification.

The basic idea in the analysis is to compute the "distance" between subjects on the basis of their categorization behavior, seeing whether the average distance between an experimental subject and the members of the control group is greater than the average distance of control group members from each other. We compute the distance between two subjects j and k for experimental item i as the number of topical categories where the subjects disagreed for this item, i.e. 0 if they placed item i into the same category and 2 if they did not (Hripcsak et al. included the more general case of allowing an item to be placed into multiple categories, so their distance measure could range from 0 to 6). The overall distance from subject j to subject k is then just their average distance across all N items. The principal question is how much the categorization behavior of subjects in the experimental (gloss translation cards) condition differs from behavior of subjects in the control (English cards) condition, computed as the average of the distance from each gloss translation card subject to every English card subject.

The upper bound on performance in this task, for subjects in the experimental group, would be categorizing the test items in just the same way as the control group categorized their items in English; this would indicate that the gloss translations were effectively providing the same information with respect to this task. The corresponding average distance for each English card subject to every other English card subject is computed in the same way as the gloss translation card subject to English card subject distance, though naturally in this case the averaging excludes the distance of each subject from himself or herself. In order to approximate lower bounds on categorization accuracy (and thus upper bounds on the average distance measure), we did eight runs placing the cards with gloss translations of Japanese into the seven categories (including "none of the above") at random. As an alternative lower bound, we also categorized the test items into the six categories (excluding "none of the above") using a forced-choice naive Bayes classifier.[4]

Figure 5 shows, for each subject, a point (and 95% confidence interval), representing its distance on average from the judgments of the subjects in the English-card (control) condition. (Recall that distances range from 0 to 2.) As one should expect, the categorization behavior of subjects given degraded information (gloss translations of Japanese) is far closer to the control group than the baselines, but generally appears to differ from that of subjects in the control group, who were given full information in the form of English cards.

Figure 6 shows the results of a second experiment, replicating the first, this time running

[4] We used McCallum's Rainbow system (companion software for Tom Mitchell, *Machine Learning*, McGraw Hill, 1997), available at ⟨http://www.cs.cmu.edu/afs/cs.cmu.edu/user/mitchell/ftp/ml-examples.html⟩.

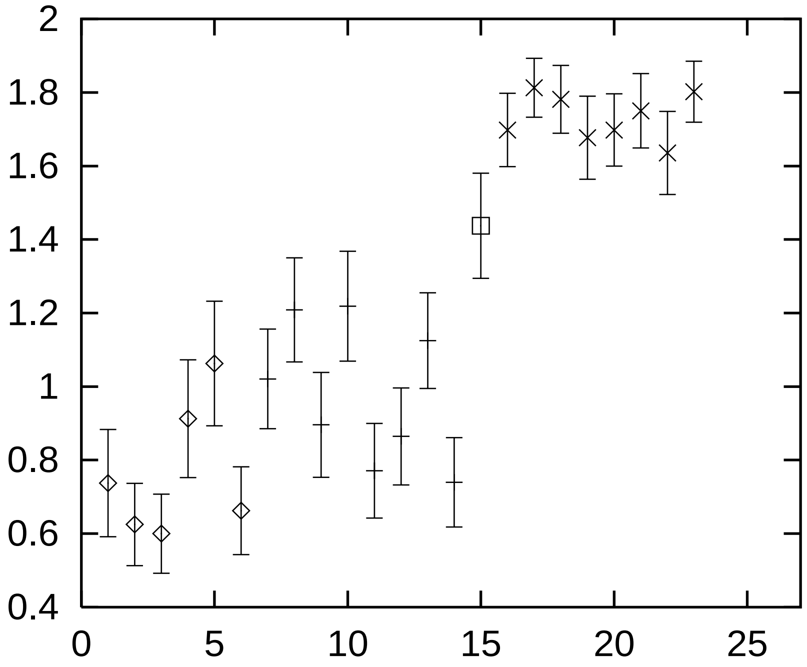


Figure 5: Results of categorization study. Left to right: Control condition (English, points 1–6), Experimental condition (translated, points 7–14), Naive Bayes classification (point 15), random classification (points 16–23). The y axis measures distance on average from participants in the control condition.

a set of 20 subjects simultaneously in a computing laboratory using on-paper materials rather than index cards, and a Web-based multiple choice form input for selecting the category for each item.[5] In this experiment we placed more subjects in the control group in an effort to determine whether the degree of variance in the control group, suggested by comparatively greater distances for the 4th and 5th points in Figure 5, would turn out to be a pattern over a larger sample. The results show that it is not: control subjects are consistently within a range of approximately 0.6–0.8 and the results for the experimental condition are consistently in the approximate range of 0.8–1.2. Participant 13 should clearly be excluded as an outlier: that participant had a visibly difficult time understanding and executing the instructions, and took more than 3 standard deviations above the mean time to complete the task.

[5] Participants in the second experiment were student and staff volunteers at the University of Maryland.

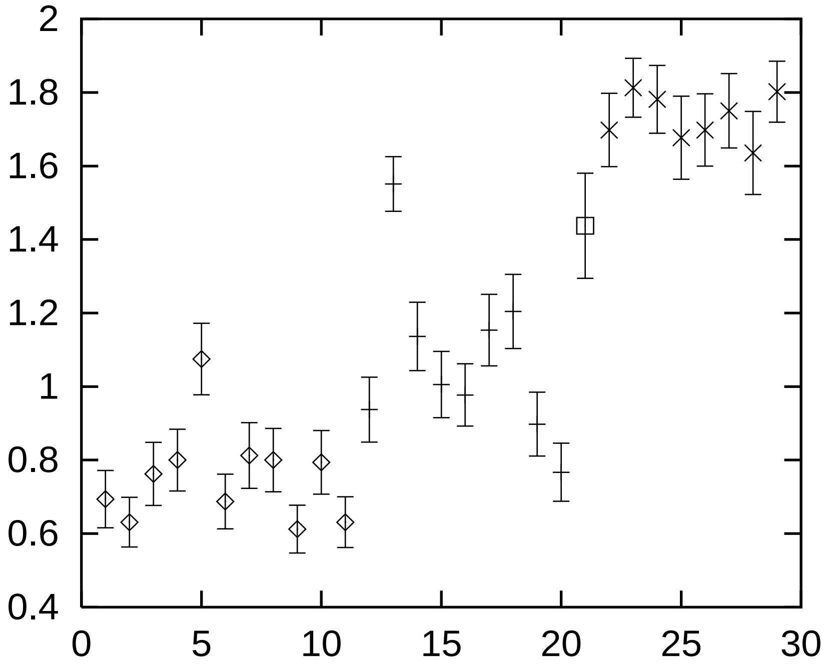


Figure 6: Results of replication. Left to right: Control condition (English, points 1–11), Experimental condition (translated, points 12–20), Naive Bayes classification (point 21), random classification (points 22–29). The y axis measures distance on average from participants in the control condition. Participant 13 is excluded as an outlier.

One advantage of the online data collection in the second study is that reliable timing data to the nearest minute were easily obtained from file modification time stamps. The results (excluding the outlier) show that although on average participants reading gloss translations took a little longer at the classification task, the difference is not significant (mean 10.7 vs 12.1 minutes total, t(17) = 1.64, p > .1).

Discussion

Taken together, the timing and performance results in the experiments suggest that simple gloss translations are useful for making categorization decisions representative of information seeking behavior at the selection interface of a cross-language information retrieval system.

Participants presented with gloss translations sorted items into categories in essentially the same amount of time as control participants given English versions of the same items, and their categorization behavior clearly resembles the categorization behavior of the control group. At the same time, the behavior of the two groups is also clearly separable, indicating that there is room for improvement in the translation technology.

In order to assess the broader utility of this experimental design for modeling real-world decisions, such as those at the selection interface, it is useful to have a clearer idea of what the categories used in this particular experiment were and what the user categorization decisions were like. In this section, we discuss the categories used in the experiment from the perspectives of *coherence* and *distinguishability*.

As noted in the previous section, the categories were obtained by asking three subjects to manually cluster cards into between four and six "piles," or categories. The descriptive labels they provided, shown in Table 1, give an idea of the distinctions they were drawing. Their categories are generally related to, but not identical to, the "Yellow Pages" categories from which the materials were taken.[6] In both the original experiment and its replication, we used the clusters given by Participant 3 as categories (the "topical category exemplars", in Figure 4), on the basis of both their number and their reasonableness. Intuitively we also found this to be the most discriminable set of the three: Participant 1 aimed at a very high level of abstraction and the first two categories are hard to distinguish, at least on the basis of the label, and Participant 2 indicated by the choice of descriptive labels that at least one cluster (Misc) lacked

[6] Since advertisements were sampled at random in order to avoid experimenter bias, one category of advertisements reflects the fact that sex-related sites are common on the Web. The Nihongo Yellow Pages directory entries used in the experiments contained only general descriptions of the information available from each of the listed services, and all experimental participants were adults.

Participant 1	Participant 2	Participant 3
Business and Personal Services Business Services Products and Catalogs Information, Organizations, and Educational Institutions	Legal/Financial Services Sex Business Appeals to Gullible People Vacation/Travel Svcs Training Materials Music/Video Sales/Sampling Misc (Service, Equipment, Business Opportunity)	Medical and Health Care Sex-Related Leisure and Recreation and Arts Recreation and Travel, especially Financial Services Personal Improvement and Education

Table 1: Category labels provided for alternative manual clusterings of experimental materials

coherence.

Qualitatively, the coherence of the selected categories is confirmed by consistency in the descriptions assigned by participants in the second part of the experiment. As described in the previous section, experimental participants were asked first to read the cards in each category, and then to write down a description for each category, to be used as a memory aid in the categorization task. The descriptions people assign, as illustrated in Figure 7, show the degree of consistency in describing the six categories. We interpret this as confirmation that participants in the categorization task are forming clear and coherent mental characterizations of the six categories that are consistent across individuals.

In addition to this qualitative assessment of category coherence, we analyzed distinguishability of the experimental categories by computing the intercategory *degree of distinguishability* for the second study on the basis of subjects' categorization decisions. Following Bruce and Wiebe (Bruce & Wiebe, 1998), for any two subjects in the categorization task we compute the degree of distinguishability for categories i and k as

$$\delta_{ij} = 1 - \frac{\hat{p}_{ij} \times \hat{p}_{ji}}{\hat{p}_{ii} \times \hat{p}_{jj}},$$

where $\hat{p}_{ij}$ is an estimate of the probability that an item placed in category i by one judge is placed

1. **Medical and Health Care** (6): Alternative/Traditional Health Resources; Medical; Medical Related items; Medical Products; Pharmaceuticals/Alternative/Oriental Medicine; Medical/Health Products; Health (self-help); Alternative Medicine Solution Suppliers; Health Care and Maintenence; Medical Supplier and Reference Site; Medical Products Services; Medical Services/Training/Supplies; Medical; Alternative/Medicine
2. **Sex-Related** (3): Erotica/Interactive; Porn; Adult Entertianment; Porno Stuff; Sex/Entertainment; X-rated Publications/Services; Erotica; Pornographic Products/Services; Erotica Site; Adult Audio/Visual Material; Sex Videos; Sex Stuff; Sexual Explicit; Pornography/Phone Sex/On-Line Sex
3. **Leisure and Recreation and Arts** (6): Music/Japanese Items; Arts; Entertainment/Hobbies; Music/Video Sutff; Music/Tourism; Cultural Products/Services; Art & Music (trade?); Arts/Entertainment Products/Services; Multi-media Audio/Visual; Music Info Site (japanese seem out of place); Multimedia Products/Services; International Arts/Bought/Sold/Displayed; Arts/Crafts/Music/Video; Arts/Entertainment
4. **Recreation and Travel, especially** (4): Travel/Immigration; Travel/Immigration/Local Spots; Travel; Travel; Travel; Travel/Vacation/Destination Services; Travel; Travel; Travel; Pre-Travel Needs and Arrangements; Travel/Immigration/Vacation Info; Travel Info/Services; Travel/Tourism; Travel/Immigration
5. **Financial Services** (5): Banking/Financial Tax Advice/Services Offshore Japanese; Finances; Financal Services; Bank/Finance Services; Tax Services; Financial Services; Ecommerce; Financial Services; Financial; Financial Services and Info; Financial Services; Financial Services; Business; Personal Finance/Investing
6. **Personal Improvement and Education** (7): Self-help/Media-based/Training/; Learning/Training; Training; Multi-media Training; Education/Technology; Business/Professional Training/Business Services; Education (safety); Education/Training Services; Education/Training; Education and Training Aids/Providers; Training/Learning Services/Products; Services to Enhance Productivity in Business; Education/Training; Education/Training

Figure 7: Descriptions assigned by participants performing the categorization task in the first study, after reading the cards in each category. The description in bold type is the label assigned by the creator of the categories, followed by the number of cards in the category.

in category j by the other. If $\delta_{ij} = 1$, the categories are said to be completely distinguishable, and if $\delta_{ij} = 0$ they are said to be completely indistinguishable.

Analyzing the categorization behavior of control subjects in order to examine distinguishability of categories, an interesting observation resulted. It became immediately clear that one of the subjects, identified as CF1, had a pattern of distinguishability values that was visibly very different from all the other subjects in the control group. For example, categories 3 and 6 are highly distinguishable — i.e., $\delta_{3,6}$ is close to 1 — for virtually all pairs of subjects. But in computations of $\delta_{3,6}$ involving categorization decisions by CF1, $\delta_{3,6} \ll 1$ in virtually every case.[7] Not surprisingly, the visible outlier in the control group, data point 5 in Figure 6, turns out to be subject CF1.

Excluding CF1, the computation of δ_{ij} strongly confirms the reliable distinguishability of pairs of categories as judged by pairs of judges. Except for the pairing of Category 2 with Category 4, to be explained momentarily, the average degree of distinguishability across pairs of judges and pairs of categories is .95 (sd = .38). It was not possible to obtain reliable values for $\delta_{2,4}$ for most pairs of judges because those categories were too sparsely populated in the set of cards to be categorized — all but one control subject left Category 2 entirely empty.

Figure 8 illustrates the small number of cards categorized by the control group into Category 2 or 4, showing, for each category, the average number of cards placed in that category for each of the conditions. This way of looking at categorization behavior provides useful information about differences between the control and experimental groups: given only limited word-by-

[7] As defined, degree of distinguishability does permit $\delta_{ij} < 0$, and we found that in a small number of cases. Bruce (personal communication) notes that δ_{ij} is guaranteed to range from 0 to 1 under certain assumptions about the distribution (Darroch & McCloud, 1986); these require more sophisticated techniques for estimating $\hat{p}$.

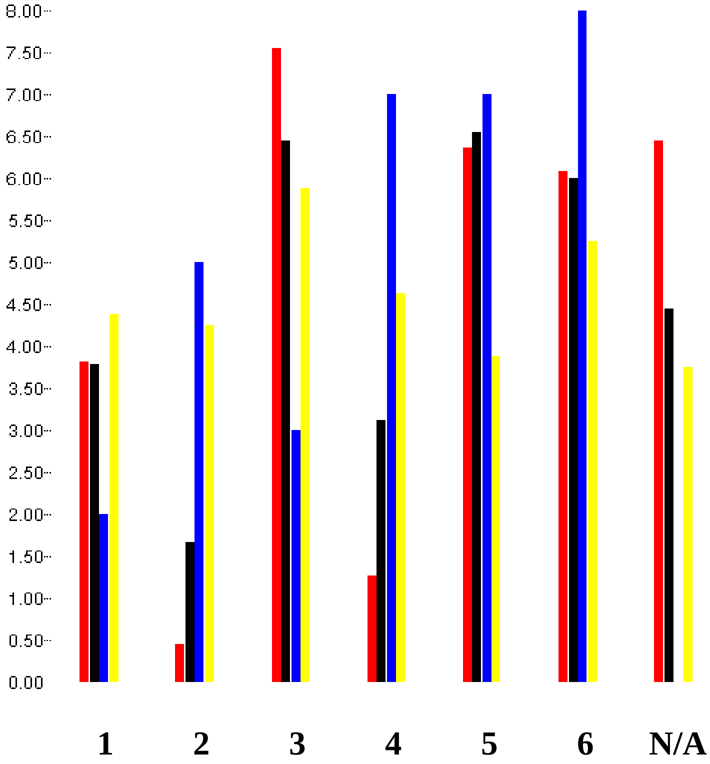


Figure 8: Average number of cards assigned to each category by (from left to right): control group (English), experimental group (translated), Naive Bayes classifier, and random selection.

word translations, many with gaps of vocabulary coverage, subjects in the experimental group placed cards into categories 2 and 4 much more frequently than subjects in the control condition. Conversely, subjects in the control group, having access to full information, used the "none of the above" category more frequently than subjects in the experimental condition.

These average frequency differences provide useful information about the way alternative presentations of information affected the decision-making process: subjects in the experimental group appear to have been misled by limited information (or by lack of confidence in their assessment) into producing false positives, attempting to categorize into one of the preexisting categories (1–6) items that control subjects viewed as belonging to the "none of the above" category.

Our analysis has also exposed three limitations of the methodology and suggested some directions for improvement. First, the effort to minimize experimenter bias through random choice led to a set of categories substantial variations in the number of items to be placed in each category, and in particular included categories with few or no test items. This makes it more difficult to draw inferences from user behavior; for example, we do not know whether misclassifications into Categories 2 and 4 reflected differing (but firm) assessments of the degree of similarity between the test items and those categories or whether it reflected an interaction between increased uncertainty about the proper category for some items and a possible bias against leaving any categories empty. Questionnaires, think-aloud protocols, and structured post-test interviews could be used to qualitatively assess the effect of such factors, and stratified sampling could perhaps be used to correct difficulties that are uncovered.

Second, we have not identified which aspects of our gloss translations are responsible for the observed differences in categorization behavior. For example, we do not yet know which possible enhancements — better vocabulary coverage, resolution of semantic ambiguity, more appropriate choice of word order, etc. — offer the greatest potential for improved utility in categorization (and hence selection) tasks. On the other hand, our experimental paradigm makes it relatively easy to vary parameters of that kind; one could easily run a new experimental condition with the same gloss translation algorithm but a better dictionary, in order to assess the role of vocabulary gaps. for example. In fact, if a similar set of subjects is available, new conditions could be added without repeating the conditions for which data has already been collected.

Finally, we have not yet evaluated the effect of gloss translations on category learning. Our present methodology allows us to focus on categorization behavior by starting from a point at which both control and experimental subjects have learned the same categories from the same six sets of English cards. Our methodology could be used to indirectly evaluate the effect of gloss translations on category learning by asking both the control and experimental groups to categorize gloss translations. The control group would learn the categories as before, using English cards, while the experimental group would learn the categories using gloss translations. Differences in the categorization behavior of the two groups could then be attributed to differences in their category learning behavior.

Conclusions

The methodology described above makes it possible to evaluate the way presentation of information — here, English versus gloss translations — influences a decision-making process — document selection in cross-language information retrieval. By adopting a categorization paradigm, and varying only the form in which information is presented, it becomes possible to augment the qualitative assessments of categorization behavior with quantitative measures such as inter-rater distance, category distinguishability, and average frequency distributions. The methodology blends the controlled conditions normally associated with highly structured evaluations with the user orientation of operational user studies. Our experiment design requires little in the way of specialized apparatus, preparation, and the like, thus facilitating replicability.

Our gloss translation strategy might be improved, for example, by biasing the presentation order of translated terms using corpus statistics, or by displaying alternative translations vertically rather than horizontally. The value added by any such changes can be measured using the same methodology by simply adding an additional condition in which subjects used the putatively improved interface. Furthermore, the technique can easily be adapted to measure the effect of other kinds of information presentation on selection decisions. For example, title translation techniques could be compared with automatically generated one-line translated document summaries to assess the utility of available summarization technology for this purpose. The methodology could also be applied in a monolingual setting, for example comparing automatic and manual techniques for constructing indicative abstracts (document summaries that are intended to support selection decisions).

Interactive information systems that effectively support end-user searching will be an important component of future digital libraries. We have shown that relatively simple interface designs can support extension of those capabilities to monolingual users in multilingual environments. While such users will undoubtedly require access to some high-quality translations, our results indicate that easily generated gloss translations may suffice to support their initial selection decisions. That could, in turn, allow designers of search technology to effectively exploit relatively efficient query translation techniques, rather than the more resource-intensive techniques based on massive document translation that are now commonly used when support for interactive selection by monolingual users is required. When combined with similar experimental methods for interfaces that support cross-language query formulation and with well known techniques for characterizing the performance of alternative search technologies, our methodology can provide a sound basis for allocating functionality in systems that are designed to support interactive information seeking in multilingual environments.

Acknowledgments

We would like to thank Gary Adams and Mona Diab for their help with the design and implementation of these experiments, Yoshi Okamoto for his assistance with Japanese figures, Rebecca Bruce and Janyce Wiebe for discussion of inter-category distinguishability, Andrew McCallum for the use of Rainbow, and Nihongo Yellow Pages Inc. for the use of their materials. This work was supported by Sun Microsystems, the Defense Advanced Research Projects Agency under DARPA/ITO Contract N66001-97-C-8540, the Department of Defense under contract MDA90496C1250, and the University of Maryland General Research Board.

References

Arnold, D., Sadler, L., & Humphreys, R. L. (1997). Evaluation: An assessment. *Machine Translation*, *8*(1–2), 1–24.

Bruce, R., & Wiebe, J. (1998). Word sense distinguishability and inter-coder agreement. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing (EMNLP-98)* Granada, Spain. Association for Computational Linguistics SIGDAT.

Church, K. W., & Hovy, E. H. (1993). Good applications for crummy machine translation. *Machine Translation*, *8*(4), 239–258.

Darroch, J., & McCloud, P. (1986). Category distinguishability and observer agreement. *Australian Journal of Statistics*, *28*(3), 371–388.

Davis, M., & Ogden, W. C. (1997). QUILT: Implementing a large-scale cross-language text retrieval system. In *Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 92–98 Philadelphia.

Hearst, M. A., & Pederson, J. O. (1996). Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of the 19th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pp. 76–84 Zürich.

Hripcsak, G., Friedman, C., Alderson, P., DuMouchel, W., Johnson, S., & Clayton, P. (1995). Unlocking clinical data from narrative reports: A study of natural language processing. *Annals of Internal Medicine*, *122*(9), 681–688.

Hull, D. A., & Oard, D. W. (1997). Symposium on cross-language text and speech retrieval. Tech. rep. SS-97-04, American Association for Artificial Intelligence, Menlo Park, CA.

Kikui, G., Hayashi, Y., & Suzaki, S. (1996). Cross-lingual information retrieval on the WWW. In *Electronic Proceedings of the Workshop on Multilinguality in Software Engineering:*

The AI Contribution (MULSAIC 96). European Conference on Artificial Intelligence. http://www.iit.nrcps.ariadne-t.gr/∼costass/mulsaic.html.

Oard, D. W., & Hackett, P. G. (1997). Document translation for cross-language text retrieval at the University of Maryland. In *The Sixth Text REtrieval Conference (TREC-6)*. National Institutes of Standards and Technology. http://trec.nist.gov/.

Resnik, P. (1997). Evaluating multilingual gisting of web pages. Tech. rep. CS-TR-3783, University of Maryland, College Park.

Taylor, R. S. (1962). The process of asking questions. *American Documentation*, *13*(4), 391–396.

Wilbur, W. J. (1994). Non-parametric significance tests of retrieval performance comparisons. *Journal of Information Science*, *20*(4), 270–284.

Additional references on cross-language information retrieval can be found at http://www.clis.umd.edu/dlrg/clir/.

About the Authors

DOUGLAS OARD is an assistant professor at the University of Maryland in the College of Library and Information Services and a member of the Digital Library Research Group. **Author's Present Address:** College of Library and Information Services, 4121G Hornbake (South Wing), University of Maryland, College Park, MD 20742. email: oard@glue.umd.edu

PHILIP RESNIK is an assistant professor at the University of Maryland, in the Department of Linguistics and the Institute for Advanced Computer Studies. **Author's Present**

Address: Department of Linguistics, 1401 Marie Mount Hall, University of Maryland, College Park, MD 20742. email: resnik@umiacs.umd.edu