Transcending the Tower of Babel: Supporting Access to Multilingual Information with Cross-Language Information Retrieval

Douglas W. Oard
College of Information Studies and Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742 USA

1. Introduction

With the advent of location-independent access to massive collections of searchable content on the World Wide Web and the convergence of text, images, audio and video in multimedia computing environments, we have come a long way towards seamless access to the information needed for commerce, security, and society. Language, however, has the potential to balkanize the information space. This chapter describes what we now know about the design of search systems that can be used to find information regardless of the language in which that information is expressed.

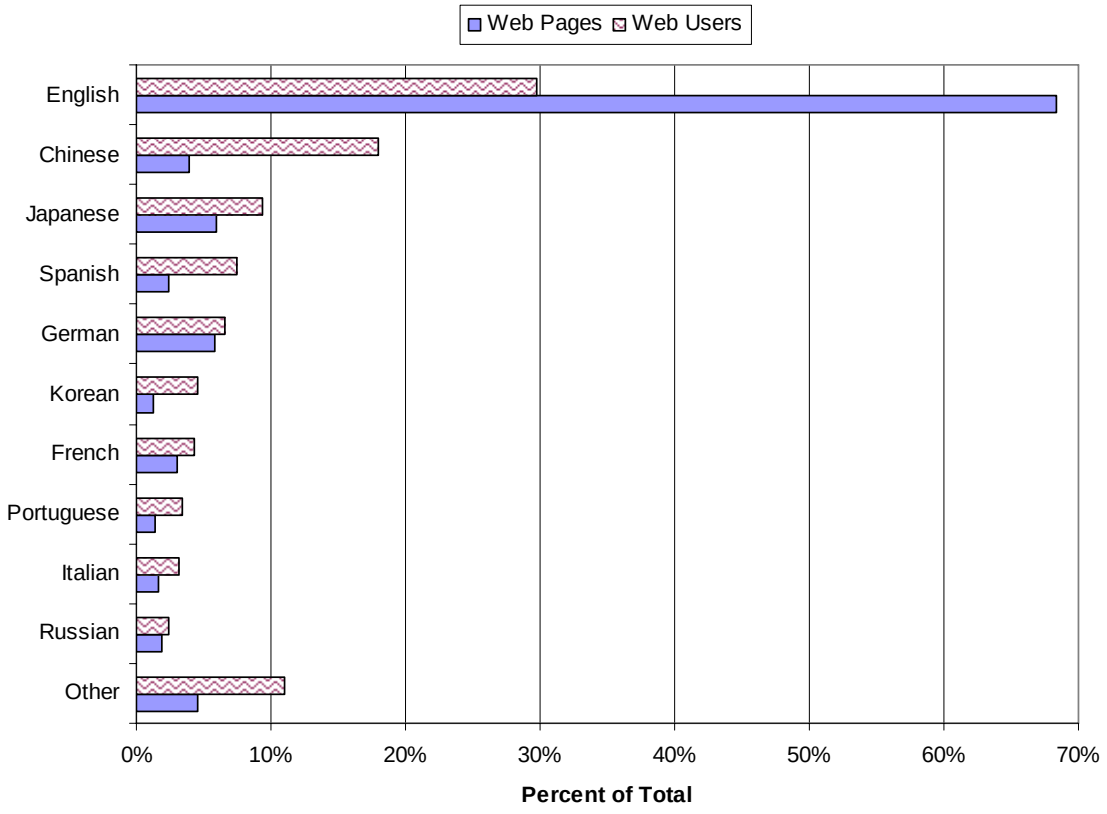

Figure 1. Language distribution of Web pages and the first language of Web users. Source: Global Reach, September, 2004

Figure 1 illustrates the nature of the challenge. The outer ring depicts the estimated fraction of the 943 million Web users that speak each of the world's languages as their first language; the inner ring depicts the estimated fraction of the Web content that is available in those languages. It is clear that English is the dominant language of Web content. Indeed, the disparity is even sharper than it first appears, since some of the non-English content is also available in English. Of course, many people, particularly those already using the Web, have a good command of English as a second language.

At each point in history, some language has dominated commercial and intellectual pursuits in the Western world, from Greek to Latin to German, and now English. So it is not surprising to see that some degree of distributional disparity between content and the first language of Web users. Indeed, looking to the future it seems reasonable to expect that situation to persist. Growth rates for speakers depend on both the fraction of speakers of that language that are presently online and on the viability of economic models that might extend Internet services to a larger portion of the population. In the near term, Chinese is the one language in which those factors come together to predict explosive growth, with much of the potential increase being among people who speak only Chinese. This will naturally lead to increased production of Chinese content, of course. But those increases may well be dwarfed by the continuing explosion of content in the other major languages of the industrialized world, and English in particular, for which large and wealthy markets already exist. So a thumbnail sketch of the near future would predict a significantly greater fraction of Chinese speakers that may well not be matched by proportional growth in Chinese content.

That brief review establishes the first major market for access to multilingual information access: Web search, for the two-thirds of Internet users for whom English is not their first language. And that market is growing. There are, however, two other obvious markets for CLIR: marketing products, and information management for national security and law enforcement operations (referred to below generically as "security" applications). The application to marketing is fairly straightforward; speakers of English presently possess the majority of the world's wealth, so producers in every region will naturally want information about their products to be easily available to English speakers.

Security applications are the most challenging scenario for CLIR because of language diversity. Estimates vary, but there are probably about 2,000 languages in common use in the world today. The public library in the New York City borough of Queens collects materials in more than 80 languages, an observation that offers some indication of the linguistic diversity with which public safety professionals must routinely cope in some major urban areas. Military operations pose even more severe challenges, both for coordination with coalition forces and for defensive or offensive information operations. The Defense Language Institute (DLI) presently trains military personnel in 31 languages for which operational needs are predictable, 13 of which were added only after the September 2001 attacks. Developing operationally significant capabilities in this way can take years, however, and our ability to predict the next flashpoint has proven to be limited. For example, DLI does not presently teach any of the four major languages spoken in Albania, but more than 5,000 soldiers were required to deploy to that country on less than 30 days notice in 1999. In that case, deployment to Macedonia was originally considered, Albania was selected on March 29 after Macedonia declined to sanction the deployment, the decision to deploy to Albania was made on April 3, and operations there began on April 23 (Nardulli, 2002). Military forces must plan for the worst case, and with thousands of languages in the world, we simply must rely on technology to augment whatever capabilities our forces are able to bring to the fight.

It is useful to think about language technologies in two groups, those that help people find information ("access technologies") and those that help people make sense of what they have found (technologies to facilitate understanding). While the two groups are certainly coupled to some degree, this natural division results in substantial simplification in system design. The key reason for this is that search is a relatively well understood process, at least when the query and the documents are expressed in the same language. The remainder of this chapter is therefore focused on extending that capability to the cross-language case, a capability that is typically referred to as Cross-Language Information Retrieval (CLIR).

The chapter is organized as follows. First, present CLIR capabilities are briefly surveyed in order to establish the present state of the art. Section 3 then draws those capabilities together, presenting three deployment scenarios that together illustrate the search capabilities that are now possible. Section 4 then presents a discussion of research investment strategies, including some prognostication on near-term commercial investments, a description of additional near-term opportunities for government investment, and identification of potentially productive investments in more basic research that could transform the opportunity space. Finally, Section 5 concludes the chapter with a few observations on the fundamental limitations of CLIR technology.

2. The State of the Art

Ultimately, it is people (rather than machines) that seek information; Information Retrieval (IR) systems are therefore best thought of as tools that help people to find what they are looking for. Three key points help to define the scope of the field. First, the information that is sought must already exist; IR systems do not create information, all they do is help people to find it. Second, IR systems are generally designed to serve a broad range of specific information needs that can not be anticipated in a detailed fashion when the system is designed. Third, IR systems are often employed iteratively, with searchers examining the results of one search iteration and using what they learn to refine the way in they express their information needs. Figure 2 illustrates one common design for CLIR systems.

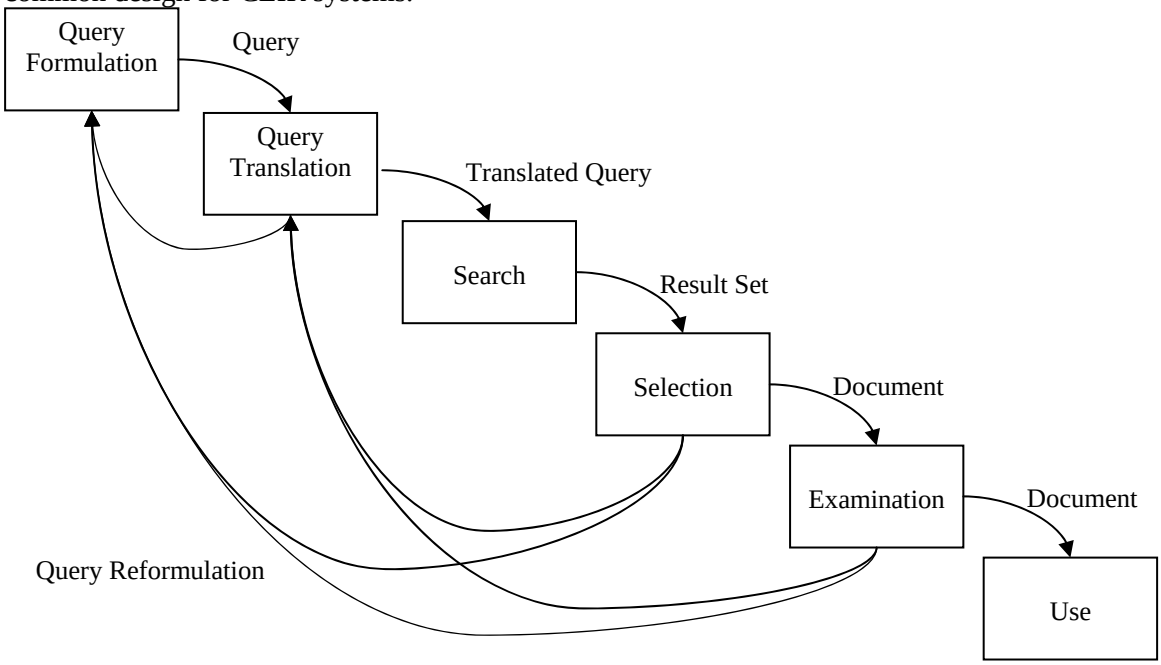

Figure 2. Component interaction for a CLIR system based on query translation.

2.1. Evaluation

IR results cannot be "right" or "wrong" in the abstract; the degree of correctness depends upon the intent and needs of the searcher. Three simplifying assumptions are generally made when evaluating the effectiveness of an IR system. First, only a single iteration is evaluated; a fixed query is proffered and the search component produces a result set. Second, only the degree to which a document is on topic ("topical relevance") is assessed; this ignores factors such as authoritativeness, degree of reading difficulty, or redundancy within the result set that are also important to users in many situations. Third, topical relevance is modeled as a binary variable; every document is treated as if it is either relevant or it is not (usually, a document is treated as relevant if any substantial part of the document is relevant).

These assumptions lead to an elegant and useful formulation for IR evaluation. Systems are asked to rank the entire document set in order of decreasing probability of relevance, and the user is modeled as wishing to examine some (unknown) number of relevant documents, scanning down from the top of the list until that number of relevant documents have been found, accurately recognizing each relevant document along the way. The user's satisfaction with the results is modeled as "precision," which is defined as the fraction of the documents that were examined that turned out to be relevant. Since the number of relevant documents that the user wishes to find is not known, a uniform distribution is assumed, and an expectation (i.e., an average value) is computed. Formally, "uninterpolated average precision" for a topic is defined as the expected value (over the set of relevant documents) of the precision at the point each relevant document appears in the list. Some systems are better for one topic than another, and we do not know in advance what topic the user will ask about. This is addressed by repeating the process for several randomly selected topics (typically, 40 or more) and reporting the expectation across the topic set of the uninterpolated average precision, a value commonly referred to as "mean average precision."

It would be impractical to judge the relevance of every document in a large collection, so a sampling strategy is needed. The usual strategy is to conduct purposive sampling on a topic-by-topic basis that is focused on the relevant documents for that topic. That can be done by first pooling the top-ranked documents (typically 100) from a diverse set of (typically 10 or more) IR systems and then judging only the documents in that pool; all other documents in the collection are then treated as not relevant. Relevance judgments formed in this way may be somewhat incomplete, but they are unbiased with respect to the systems that contributed to the pools. Importantly, hold-one-out studies have shown that judgment sets constructed in this way are also unbiased with respect to other IR systems of similar design (Voorhees, 2000). Although people sometimes disagree about the topical relevance of individual documents, multi-judge studies have shown that replacing one judge's opinions with another's rarely changes the preference order between systems. Evaluations using judgments reported in this way are therefore best reported as comparisons between contrastive system designs rather than as absolute measures of effectiveness, since different users may assess the relevance of retrieved document differently. For CLIR experiments, the reference value is typically the mean average precision achieved by a system of comparable design using queries in the same language as the documents (a "monolingual baseline").

With that as background, it is now possible to describe the effect of the known CLIR techniques in terms of this evaluation framework. Large CLIR test collections (often with more than 100,000 documents) are presently available with documents (typically, news stories) in Arabic, Bulgarian, Chinese, Dutch, English, Finish, French, German, Hungarian, Italian, Japanese,

Korean, Portuguese, Russian, Spanish, and Swedish, and the results reported here are generally typical of what is seen for those languages (Braschler, 2004; Kishida, 2004; Oard, 2002).

2.2. Techniques

The basic strategy for building any IR system is to represent the documents in some way, to represent the query is some compatible way, and then to compute a score for each document using a function of the query representation and the document representation that (hopefully) assigns higher values to documents that are more likely to be relevant. Counting terms (where terms may be parts of words, full words, or sequences of words) has proven to be a remarkably useful basis for computing document representations. Three factors are typically computed: (1) term frequency (TF), the number of occurrences of a term in a document; (2) document frequency (DF), the number of documents in which a term appears; and (3) length, the total number of terms in a document. Essentially, DF is a measure of term selectivity, while the ratio between TF and length is a measure of aboutness. These factors are used to compute a weight for each term in each document, with higher TF, lower DF, and shorter length resulting in higher weights. The most effective weighting functions (e.g., Okapi BM 25) also typically transform the TF and DF factors in ways that grow more slowly than linear functions, and some systems also factor in additional sources of evidence (e.g., term proximity). The score for each document is then computed (at query time) as the sum of the weights of the query terms in that document. The documents can then be sorted in decreasing score order for presentation to the user.

CLIR applications introduce one obvious complication: the query and the documents contain terms from different languages, so direct lexical matching will often not be possible. Three basic approaches to overcoming this challenge are possible: (1) map the document language terms into the query language, (2) map the query language terms into the document language, or (3) map both document language and query language terms into some language-neutral representation. Because each term is processed independently in a typical IR system, these mappings are typically done on a term-by-term basis. Term translation poses three challenges for system design: (a) selection of appropriate terms to translate, (b) identifying appropriate translations for each term, and (c) effectively using that translation knowledge.

Three sources of translation mappings are available to an automated system: (1) a bilingual or multilingual lexicon, (2) a bilingual or multilingual corpus, and (3) sub-word translation mapping algorithms. While all three are useful, the most effective systems rely on bilingual “parallel” corpora that contain documents written in one language and (human-prepared) translations of those documents into the other. Through automatic sentence alignment, term selection, and within-sentence term alignment, it is possible to compute not just the possible translations for a term, but also to estimate the probability that each possible translation would be used. Figure 3 shows one possible set of alignments for the first few Spanish and English words from a parallel corpus of Spanish and English proceedings of the European Parliament. For probabilities estimated in this way to be most useful, the parallel text collection should be large (so that the translation probabilities can be accurately estimated) and it should use language in a manner that is similar to they way language is used in the documents to be searched (e.g., it should be from a similar genre, with similar topical coverage). Suitable parallel text collections can often be found, since the same factors that lead to a need for CLIR typically also result in manual translation of at least some materials that are in particularly high demand (Resnik, 2003). When that is not the case, focused elicitation of the needed translations is sometimes a viable alternative (Yarowsky, 2003).

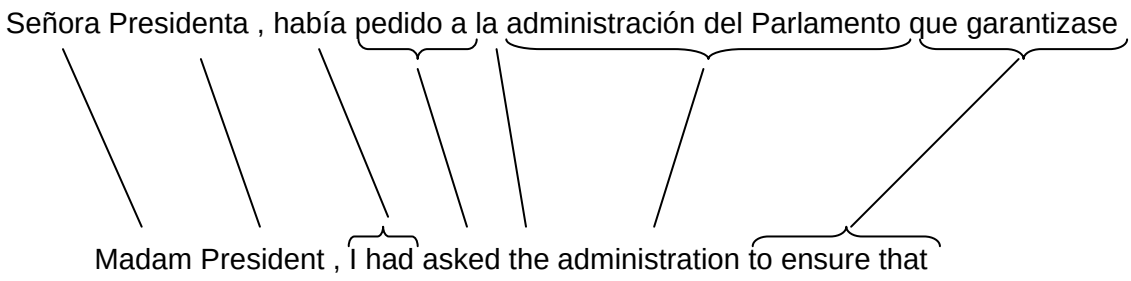

Figure 3. An example of aligning Spanish and English terms (source: EUROPARL corpus).

Although translation probabilities extracted from parallel text are quite useful, there are two cases in which parallel texts do not yield useful translations: (1) uncommon terms, which may not appear sufficiently often in even very large parallel text collections, and (2) terms that were introduced after the parallel text was collected. Hand-built translation lexicons (e.g., bilingual dictionaries) can be a reliable source of translation knowledge for the first case; a uniform distribution on translation probability can be assumed if no information about preferred translations is supplied with the lexicon. Newly coined terms may be missing from translation lexicons, but it is sometimes possible to predict the way in which such a term will be translated by mapping pieces of the term separately and then reassembling the translated pieces. For example, names of people are often translated from English into Chinese by sounding out the English word and then selecting Chinese characters that would be pronounced similarly (Lin, 2002). A similar approach can be used to translate multiword expressions; every known translation of each constituent word is postulated, and then a large collection of text in the target language is used to select the one combination that occurs together most often (López-Ostenero 2001).

With adequate translation knowledge in hand, the translation process itself is quite straightforward. One good approach is to separately map the TF and DF evidence, allocating weight across the known translations using the estimated translation probabilities (Darwish, 2003). This approach can be applied to map in either direction, and it can be helpful to merge evidence from both translation directions because the error characteristics of the two mappings are often complementary (McCarley, 1999). When well integrated, it is possible to exceed 100% of a credible monolingual baseline system's mean average precision using these techniques. It may seem surprising at first that any cross-language technique could exceed a monolingual baseline, but this merely points up a limitation of comparative evaluation; it is difficult to introduce synonyms in monolingual systems in a manner that is comparable to the synonyms that are naturally introduced as a byproduct of translation, so (relatively weak) monolingual baselines that lack synonym expansion are often reported.

Table 2 shows a typical example of CLIR results based on a translation model learned from parallel European Parliament proceedings in English and French. The results for the unidirectional case were obtained using test collections developed for the Cross-Language Evaluation Forum (CLEF) using the following formulae from (Darwish, 2003):

$$TF_d(e) = \sum_{f_i \in T(e)} p(f_i \mid e) \times TF_d(f_i)$$

$$DF(e) = \sum_{f_i \in T(e)} p(f_i \mid e) \times DF(f_i)$$

$$w_d(e) = \frac{TF_d(e)}{1.5\frac{L_d}{L_{avg}} + TF_d(e) + 0.5} \times \log\left(\frac{N - DF(e) + 0.5}{DF(e) + 0.5}\right)$$

$$s_d = \frac{1}{|Q|}\sum_{e \in Q} w_d(e)$$

where the symbols have the meanings shown in Table 1. For the unidirectional case, p(fi|e) was estimated using the freely-available Giza++ software using English as the source language and French as the target language. For the bidirectional case, the Giza++ system was run with the source and target languages swapped, the results were inverted using Bayes rule, and the results of that reversal were averaged with the function learned for the original translation order. This is now a standard technique in the design of machine translation systems because it helps to compensate for an asymmetry that Giza++ and similar systems introduce for efficiency reasons. Finally, for the monolingual case the first two equations were not needed and f as use in place of e (i.e., French queries were used).

Symbol	**Meaning**
S_d	Score for document *d*, the basis for ranking in decreasing order
Q	The set of English query terms chosen by the user
$w_d(e)$	Weight for English query term *e* in document *d*
$TF_d(f_i)$	The number of times French term f_i occurs in document *d*
$TF_d(e)$	Estimated number of times English term e would have occurred in a translation of *d*
$DF(f_i)$	The number of documents in which French term f_i occurs
$DF(e)$	Estimated number of translated documents that would have contained English term *e*
N	The number of documents in the collection
$T(e)$	The set of known French translations for English query term *e*
$p(f_i\|e)$	The probability that English term e translates to French term f_i
L_d	The number of French terms in document *d*
L_{avg}	The average number of French terms in a document

Table 1. Factors that affect the score of a document.

Table 2 shows the results. The monolingual and bidirectional CLIR conditions were statistically indistinguishable (by a Wilcoxon signed rank test for paired samples); the retrieval effectiveness of the unidirectional CLIR condition was found to be significantly below that of the monolingual condition (by the same test, at $p<0.05$). From this we conclude that training Giza++ in both directions is helpful, and that in this task the retrieval effectiveness in the monolingual and cross-language conditions are comparable. Note that this was achieved using parallel text from a different domain from the documents being searched, indicating that this technique is reasonably robust.

	Mean Average Precision
Monolingual	0.3856
Unidirectional CLIR	0.3714
Bi-directional CLIR	0.3780

Table 2. Mean (over 151 topics) average precision for monolingual search, CLIR with English queries trained in one direction, and CLIR with English queries trained bidirectionally, searching 87,191 French news stories, scored using CLEF relevance judgments.

An alternative approach that does not require parallel text is the Generalized Vector Space Model (GVSM), in which each term is represented as a vector in which each element is the frequency of the term of interest in one "training document;" the length of such a vector is the number of documents in a collection. If each bilingual document is formed by conjoining a pair of comparable documents (i.e., separately authored documents writing about the same subject, one in each language), the resulting vector space will be language-neutral. A document in the collection to be indexed (or a query) can then be represented in the GVSM vector space as the sum of the vectors for each term that it contains. Further improvements can often be obtained by applying a dimensionality reduction technique (e.g., singular value decomposition) to the matrix of term vectors before computing the representations of the documents and the query; this approach is known as Latent Semantic Indexing (LSI). Because GVSM and LSI are based on document-level alignment rather than word-level alignment, it is difficult to achieve levels of effectiveness that are competitive with what could be attained using parallel text; around 70% of the mean average precision for a comparable monolingual baseline is typically reported. Useful comparable texts may be easier to obtain than useful parallel texts in some applications, however, particularly for language pairs across which little digital interaction is presently occurring for economic or social reasons. Pairing of comparable documents is needed before such collections can be used in this way, and techniques for that task have been demonstrated in restricted domains (Sheridan, 1998).

Comparable corpora can also be used as a basis for unsupervised adaptation of translation resources to a specific application. The basic idea is to mine a source-language collection for translatable terms that might plausibly have been included in the query (but were not), then to mine a target-language collection for terms that might plausibly have resulted from translation (but did not). The standard way of doing pre-translation "expansion" is to identify documents that are similar to the query in the source language collection, and then to adjust the term weights in a way that rewards presence in the highest ranked (i.e., most similar) documents. Post-translation adaptation is accomplished in the same manner (often with the addition of term reweighting), but using a large target-language collection. Because this is an unsupervised variant of the same process that systems employ when users designate a few relevant documents for query enhancement, the process is generally known as "blind" relevance feedback. A similar approach (substituting documents for queries) can also be used with document-translation architectures. Pre-translation feedback has proven to be particularly effective when the available translation resources are relatively weak (e.g., a small translation lexicon with no access to parallel text). When pre-translation and post-translation blind relevance feedback are used together with a relatively large lexicon that lacks translation probabilities, up to 90% of the mean average precision achieved by a credible monolingual baseline (without synonym expansion) has been reported. This compares favorably to the 80% relative effectiveness that is typically reported under comparable conditions without blind relevance feedback. However, when large domain-specific parallel text collections are available, blind relevance feedback offers less potential benefit.

In summary, it is now possible to build systems that accept queries in one language and rank documents written in another nearly as well as systems for which both the queries and the documents are expressed in the same language. Moreover, a range of techniques are known for optimizing the use of different types of language resources, so it is reasonable to expect that such systems can actually be constructed for real applications. It is important to recognize that these

claims are based on averages, both over the topic of the query and over the position of a relevant document in the ranked list; results for individual queries and/or documents will naturally differ.

A glance back at Figure 2 will reveal, however, that construction of a ranked list is only one part of a complete search process. It is therefore also important to ask how well searchers can employ this capability in actual cross-language search tasks. We know from task-specific evaluations of machine translation that any reasonable translation system will often suffice to support recognition of documents that the user wishes to see (although at the cost of somewhat greater human time and effort) (Oard, 2004). The interaction between translation and summarization in the document selection stage has been less thoroughly studied, but anecdotal evidence from end-to-end search tasks indicate that simple combinations of summarization techniques developed for monolingual applications (e.g., extraction of fixed-length passages around query terms) and available machine translation systems works well enough. Query formulation is perhaps the least well understood area at present; interactions between vocabulary learning, concept learning, and query creation are complex, and the research reported to date has not yet fully characterized that design space. End-to-end user studies have, however, demonstrated that users can often iteratively formulate effective queries by manually entering text in the query language, and explanatory interfaces have started to appear that seek to help the user understand (and thereby better control) the cross-language search progress. Experiments with users in the loop are expensive, and thus relatively rare, but half a dozen teams have reported results, some over many years. As an example of what could be achieved as of 2004, an average of 70% of factual questions were answered correctly by searchers that could not read the document language and did not previously know the answer; that is about twice the fraction of correct answers that had been achieved by the best fully automatic cross-language question answering systems at that time (Gonzalo, 2004).

So machines can rank documents written in languages different from the query, and searchers can effectively exploit that capability for some real search tasks. The next section examines the implications of that capability by presenting three deployment scenarios that can be supported by present CLIR technology.

3. Near-Term Deployment Scenarios

Cross-language information retrieval has sometimes been uncharitably called “the problem of finding documents that you can’t read.” Why would someone want to do that? This section describes three scenarios in which such a capability could be useful.

Polyglots. Polyglots (people who are able to read several languages) are obvious candidates as CLIR system users for two reasons. First, some savings in time and effort might be realized if the searcher can formulate (and refine) their query just once, with the system then calling to their attention potentially relevant documents in any language that they can read. Depending on the number of languages involved, the results might best be displayed as separate ranked lists for each language or as a single merged list. The more significant reason that a polyglot may choose to use a CLIR system, however, is that their passive language skills (e.g., reading) and active language skills (e.g., query formulation) may not be equally well developed. In such cases, we can think of the CLIR system as a form of “language prosthesis” that can help them with query formulation and refinement. The Defense Language Transformation Roadmap calls for incorporation of language training as a regular part of professional development within the officer corps; when fully implemented, this policy will dramatically expand the number of polyglot users in the U.S. armed forces (DoD, 2005).

Team Searching. Complex information needs are best addressed when a nuanced understanding of what is being sought can be combined with the search skills to that are needed to get the best results from available systems and the language skills that are needed to make sense of what is found. These competencies need not all be present in a single individual, however. For example, search intermediaries (e.g., librarians) are often employed for high-stakes searches in fields such as medicine and law. A similar approach can be applied in the cross-language case, teaming a searcher that knows their own needs well with a skilled intermediary that has the necessary language skills to help the searcher understand (rather than simply find) the available information. Co-presence may not be essential when working in a networked environment; "remote reference services" have been the focus of considerable research recently, and tools for synchronous interaction that have already been developed (e.g., coupled displays, augmented with text chat) (Coffman, 2001) could be extended to support cross-language applications.

Two-stage triage. Scenarios that require sifting through large quantities of information in a less commonly taught language place a premium on maximizing the productivity of the small number of individuals that possess the needed language skills. In such cases, initial searches can be done using interactive CLIR systems by many skilled searchers who understand what is being sought. As promising documents are found, they can then be passed on to the few available language experts. The searchers and the language experts need not even work for the same organization; for example, promising documents might simply be submitted to a translation bureau (e.g., the National Virtual Translation Center) that will optimize the allocation of documents across the available pool of translators.

Each of these scenarios can be accomplished with the search and translation technology that is available today, but future improvements in translation technology could yield an even greater range of useful capabilities. The next section considers those possibilities.

4. Crafting an Investment Strategy

Yogi Berra is credited with having observed that prediction is difficult, particularly when predicting the future. But if we are to create a rational strategy for investing government resources, we should start with some idea of what is likely to happen even without that investment. Accordingly, this section begins with a brief survey of the commercial landscape and the prognosis for near-term developments in that sector. A discussion of additional near-term investment opportunities then follows. The discussion concludes with an articulation of the fundamental challenges that remain open; those are the candidates for continued investment in basic research.

4.1. Commercial Prospects

The single most consequential commercial development over the past decade has been the emergence of World Wide Web indexing as a commodity product. Commercial investments in search technology are driven by two key factors: affordability and market size. Automatic language identification and on-demand machine translation are now widely available, but none of the major search engines have integrated anything but the most rudimentary cross-language search technology. Affordability is certainly not the limiting factor in this case; efficient CLIR techniques have been known for several years. Rather, the problem seems to be that the market size is perceived to be sensitive to the availability of high-quality translation services. Present on-demand machine translation services are adequate for a limited range of uses, but their translation

quality (accuracy and fluency) leaves a lot to be desired, and the computational cost of the state-of-the-art "transfer method" machine translation approach used by present Web translation services is far larger than the computational cost of Web search. Broad commercial adoption of cross-language search is therefore limited far more by deficiencies in present machine translation technology than by any limitations of the CLIR technology itself.

Statistical machine translation is rapidly emerging as a practical alternative to the earlier "transfer method" approaches. Modern statistical translation systems offer two main advantages: (1) once a statistical system has been built for one language pair, it can be extended to additional language pairs with an order of magnitude less effort than was the case for transfer-method systems (about one person-year vs. about ten), and (2) statistical machine translation systems have demonstrated improved translation quality in some applications. Statistical machine translation faces two key limitations, however: (1) research investments have focused more on translation quality than on speed, so the older "transfer method" systems are currently generally faster, and (2) deploying a statistical system requires "training data" that is representative of the materials to which it will ultimately be applied (e.g., a statistical system trained using news stories might not do as well as a "transfer method" system when used to translate text chats). Recent press reports indicate that some commercial investment is now focused on addressing these two limitations. If those efforts are successful, we could see widespread deployment of CLIR technology in Web search engines over the next few years. Other, more specialized, applications (e.g., for libraries, patents, law, and medicine) could naturally follow from the demonstrated utility of the technology that would result.

Another scenario that could result in near-term commercial adoption of CLIR technology would be close coupling of cross-language search with translation routing technology. Translation routing systems seek to automatically optimize the assignment of documents to human translators in a way that balances cost, quality (e.g., by accounting for subject matter expertise), and timeliness. Access patterns in large collection are typically highly skewed (meaning that a few documents are read by many people, and many documents may be read by nobody). If one translation routing service were to capture a significant market share, this sharply focused reuse could be exploited by cacheing translations as they are created, thus amortizing translation costs over multiple users. The resulting balance between affordability, quality, and responsiveness, when coupled with the complementary characteristics of machine translation systems, could help to push the incentive for adoption of CLIR technology past the tipping point. Some policy issues (e.g., the treatment of cached translations under international copyright conventions) may need to be worked out before that can happen, however.

5.2. A Near-Term Government Investment Strategy

It therefore seems likely that near-term commercial investments will ultimately yield a broader experience base with the integration of CLIR technology in realistic operational scenarios, but some targeted government investments will also likely be needed if we are to exploit the full potential that this technology offers. For example, support for cross-language team searching will require a development effort for which no likely source of commercial investment can presently be identified. Investments in several more narrowly focused technical issues could also pay off handsomely in the near term (e.g., optimal support for query refinement in CLIR applications, effective techniques for merging result lists across languages, and closer integration of query-based summarization and machine translation technologies).

One important class of near-term investment opportunities that is almost certain not to attract commercial investment is urgent deployment of CLIR technology for new language pairs. As the

Albanian example at the beginning of this chapter indicates, deployment timelines for military forces are often far shorter than commercial development timelines could possibly accommodate. In 2003, the Defense Advanced Research Projects Agency (DARPA) conducted a "surprise language" exercise in which research teams were challenged to develop machine translation, CLIR, summarization, and information extraction technology for unexpected language pairs (Oard, 2003). A preliminary 10-day effort for the Cebuano language and a large-scale 29-day effort for Hindi both indicated that usable systems could be deployed far more rapidly than had previously been demonstrated. A balanced investment strategy in which optimized systems that are built in advance to meet predictable requirements are augmented with a flexible rapid-response capability could be implemented using technology that is presently in hand. Early designs of such a system could then be improved over time as experience in actual operational settings is gained. Unless we think that the world will be a much more stable and predictable place in the near future, we would be wise to pursue such a course.

5.3. Investments in Basic Research

A balanced investment strategy also calls for balance between near-term and long-term investments. Advances in machine translation technology would be very highly leveraged, making that the single most important focus for longer-term investments. Clear potential exists for substantial advances in translation quality, robustness, and speed through three promising avenues: (1) exploiting massive collections of naturally occurring training data (e.g., Resnik, 2003), (2) improved models of language based on closer coupling between statistical and symbolic techniques (e.g., Chiang, 2005), and (3) adaptation to unique needs of specific application environments (e.g., Warner, 2004). The rapid progress in the accuracy and fluency of machine translation in recent years has been a direct consequence of the widespread adoption of affordable and insightful evaluation techniques; continued refinement of those evaluation techniques will likely be an important prerequisite to future progress as well.

The vast majority of the work to date on CLIR has assumed that that the words to be found and translated are already represented in a "character-coded" form that makes digital manipulation of those words fairly straightforward. Of course, most of the words produced by the world's 6.4 billion people are spoken rather than written. Fairly accurate automatic transcription of news broadcasts has been possible for several years, and more recently there have been substantial improvements in the accuracy of automatic transcription of conversational speech as well (Byrne, 2004). Integration of that speech technology with CLIR and machine translation would therefore be a highly leveraged investment. Similarly, automatic recognition of printed characters is now quite accurate, and reasonably accurate automatic transcription of handwritten text is possible in some situations. Spoken, printed and handwritten content pose unusual challenges for interactive CLIR systems, however, because straightforward design options yield a cascade of errors (with transcription errors compounded by translation errors) (Schlesinger, 2001). Designing effective interactive CLIR systems requires that these issues be addressed, potentially in different ways, in at least four system components (query formulation, automated search, result list selection, and item-level examination). The proliferation of digital audio recording and digital image acquisition technology promises to move these issues to the forefront of the research agenda over the next several years.

Two other broad trends in information access technologies will also likely create important new opportunities for employment of CLIR technology: (1) search over conversational text, and (2) true "text mining." Much of the investment in search and translation technology has focused on carefully written content (e.g., news stories), but the explosive growth of conversational text genre such as electronic mail, instant messaging, and "chat rooms," provides a strong incentive to

understand how information access in large conversational genre collections will differ. The questions range from the most fundamental (e.g., "what will people look for?"), through many that are more sharply technical (e.g., "how should the possibility of typographical errors be accommodated?"), to some that are well beyond the scope of this chapter ("what archives of instant message conversations are likely to be available?"). Among the issues that will need to be addressed are mixed-language conversations, the use of sublanguage among conversational participants what share extensive context, and the consequences of informality (e.g., ungrammatical usage and iconic representations for emotions). Each of those factors promises to add complexity to the lexical mapping that underlies CLIR techniques that were originally developed for more formal genre.

The term "text mining" has been used to market a broad range of information access technologies (including, in marketing literature, ordinary query-based search systems). As a research challenge, however, it is often understood to refer to searching based on broad patterns (e.g., "find people that espouse positions on Kurdish autonomy that are rarely presented in the U.S. media") (Hearst, 2003). Satisfying information needs of that sort with any significant degree of automation can be a daunting challenge even when all the text is in the same language. Some progress in this direction has already been made, however. For example, the emerging field of visual analytics couples computational linguistics with information visualization to construct presentations that facilitate recognition of patterns in the use of language (Wong, 2004). Multi-document summarization systems (Schiffman, 2002) and the closely related work on systems for automatically answering complex questions (Diekema, 2003) adopt an alternative approach, selecting useful snippets of text and reshaping them into text-based products that the user can then (hopefully) read for comprehension. All of these technologies rely on computational models of meaning that are necessarily weak, since the ambiguity that is central to natural language resists precise modeling. Introducing additional languages will exacerbate that challenge, compounding ambiguity of interpretation with the ambiguity that results from imprecise translation. But the ability to reason automatically across large multilingual collections would also create important new opportunities by dramatically expanding the breadth of information sources and the diversity of perspectives that could be leveraged. Extending text mining technologies to multilingual applications will therefore likely merit significant investment in the coming decade.

5. Summary and Outlook

Useful cross-language search technology is available now, and with a small set of targeted near-term investments we would be in an excellent position to leverage that important capability. As with any transformational technology, however, we must couple our thinking about the design of systems with innovative thinking about how those systems will be used. The scenarios outlined above (enhancing search capabilities for polyglot users, forming search teams with synergistic skill sets, and two-level strategies that optimize the workload for personnel with scarce language expertise) represent a first step in that direction. But true organizational innovation requires experience, and gaining experience requires that we build systems. So spiral development strategies will be a natural part of the process by which this new technologies is adopted.

Some of the technology needed to provide access to multilingual content is now quite mature. We can, for example, match content with queries across languages about as well as we can in the same language. But effective searching demands synergy between searcher and system. Sustained investment in both basic and applied research will be needed if we are to optimize that synergy over the full range of potentially important applications. There are, of course, some

fundamental limits to what can be done. Existing term-based techniques for building ranked lists are far from perfect, but experience has shown that they are both useful in their present state and hard to improve upon; greater precision can certainly be achieved using techniques with greater linguistic sophistication, but only at some cost in coverage (i.e., recall) and flexibility. So now that we are able to search across languages as well as we do within the same language, focusing solely on building better ranked lists seems as if it would be a questionable investment. Instead, the time has come to refocus our efforts on the new opportunities that our past success has generated. We find ourselves at an inflection point now. Having developed the core technology for searching across languages, we are now presented with unprecedented opportunities to build deployable systems for at least the formal document genre that we have already mastered, while simultaneously beginning to explore more advanced techniques for searching conversational media in several languages and for exploratory mining of multilingual text collections.

Alvin Toffler tells us of a "third wave," a society in which information is the raw material, and the processes and systems that help people manage that information are the means of production (Toffler, 1980). Since time immemorial, men and women have sought the high ground to provide them with advantage as they struggle with their adversaries. In a conflict of ideas, the high ground is not to be found at the top of a hill, in the sky, or even in outer space; the high ground is the human mind. Language provides a window on the mind, and those who best command the realm of language will naturally be best advantaged in the competition of ideas. This is a challenge from which we simply can not shrink.

References

(Braschler, 2004) Martin Braschler and Carol Peters, "Cross-Language Evaluation Forum: Objectives, Results, Achievements," *Information Retrieval*, 7(1-2)7-31, 2004.

(Byrne, 2004) William Byrne, David Doermann, Martin Franz, Samuel Gustman, Jan Hajic, Douglas Oard, Michael Picheny, Josef Psutka, Bhuvana Ramabhadran, Dagobert Soergel, Todd Ward and Wei-Jing Zhu, "Automated Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives," *IEEE Transactions on Speech and Audio Processing*, 12(4)420-435, 2004.

(Chiang, 2005) David Chiang, "A Hierarchical Phrase-Based Model for Statistical Machine Translation," in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, MI, 2005.

(Coffman, 2001) Steve Coffmann, "We'll Take it from Here: Developments We'd Like to See in Virtual Reference Software," *Information Technology and Libraries*, 20(3)149-153, 2001.

(Darwish, 2003) Kareem Darwish and Douglas W. Oard, "Probabilistic Structured Query Methods," in *Proceedings of the 26th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, July, 2003.

(Diekema, 2003) Anne R. Diekema, Ozgur Yilmazel, Jiangping Chen, Sarah Harwell, Lan He and Elizabeth D. Liddy, "What do You Mean? Finding Answers to Complex Questions," in *Proceedings of the AAAI Symposium on New Directions in Question Answering*, Stanford, CA, March, 2003.

(DoD, 2005) Department of Defense, *Defense Language Translation Roadmap*. January, 2005.

(Gonzalo, 2004) Julio Gonzalo and Douglas W. Oard, "iCLEF 2004 Track Overview," in *Working Notes for the CLEF 2004 Workshop*, Bath, UK, September, 2004.

(Hearst, 1999) Marti A. Hearst, "Untangling Text Mining," in *Proceedings of the 37th Annual Conference of the Association for Computational Linguistics*, College Park, MD, June, 1999.

(Kishida, 2004) Kazuaki Kishida, Kuang-hua Chen, Sukhoon Lee, Kazuko Kuriyama, Noriko Kando, Hsin-His Chen, Sung Hyon Myaeng and Koji Eguchi, "Overview of the CLIR Task at the Fourth NTCIR Workshop," in *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies*, Tokyo, Japan, June, 2004.

(Lin, 2002) Wei-Hao Lin and Hsin-Hsi Chen, "Backward Machine Transliteration by Learning Phonetic Similarity," in *Sixth Conference on Natural Language Learning*, Taipei, Taiwan, August 2002.

(López-Ostenero, 2001) Fernando López-Ostenero, Julio Gonzalo, Anselmo Penas and Felisa Verdejo, "Noun Phrase Translations for Cross-Language Document Selection," in *Evaluation of Cross-Language Information Retrieval: Second Workshop of the Cross-Language Evaluation Forum*, Darmstadt, Germany, September, 2001.

(McCarley, 1999) J. Scott McCarley, "Should we Translate the Documents or the Queries in Cross-Language Information Retrieval?," in *27th Annual Meeting of the Association for Computational Linguistics*, College Park, MD, June, 1999.

(Nardulli, 2002) Bruce R. Nardulli, Walter L. Perry, Bruce Pirnie, John Gordon IV, and John G. McGinn, *Disjointed War, Military Operations in Kosovo*, RAND, 1999.

(Oard, 2002) Douglas W. Oard and Frederic C. Gey, "The TREC-2002 Arabic-English CLIR Track," in *The Eleventh Text Retrieval Conference (TREC-2002)*, Gaithersburg, MD, pp. 17-26, November 2002.

(Oard, 2003) Douglas W. Oard, "The Surprise Language Exercises," *ACM Transactions on Asian Language Information Processing*, 2(2)79-84, 2003.

(Oard, 2004) Douglas W. Oard, Julio Gonzalo, Mark Sanderson, Fernando López-Ostenero and Jianqiang Wang, "Interactive Cross-Language Document Selection," *Information Retrieval*, 7(1-2)205-228, 2004.

(Resnik, 2003) Philip Resnik and Noah A. Smith, "The Web as a Parallel Corpus," *Computational Linguistics*, 29(3)349-380, 2003.

(Schlesinger, 2001) C. Schlesinger, M. Holland and L. Hernandez, "Integrating OCR and Machine Translation on Non-Traditional Languages," in *Proceedings of the 2001 Symposium on Document Image Understanding Technology*, pp. 283-287, Columbia, MD, 2001.

(Schiffman, 2002) Barry Schiffman, Ani Nenkova and Kathleen McKeown, "Experiments in Multidocument Summarization," in *Proceedings of the 2002 Human Language Technology Conference*, San Diego, CA, March, 2002.

(Sheridan, 1998) Páraic Sheridan, Jean Paul Ballerini and Peter Schäuble, “Building a Large Multilingual Test Collection from Comparable News Documents,” in Gregory Grefenstette, ed., *Cross Language Information Retrieval*, Chapter 11, Kluwer Academic, 1998.

(Toffler, 1980) Alvin Toffler, *The Third Wave*, Bantam Books, 1980.
(Voorhees, 2000) Ellen M. Voorhees, “Variations in Relevance and the Measurement of Retrieval Effectiveness,” *Information Processing and Management*, 36(5)697-716, 2000.

(Warner, 2004) John Warner, Bill Ogden, and Melissa Holland, “Cross-Language Collaboration Between Distributed Partners Using Multilingual Chat Messaging,” in *SIGIR 2004 Workshop on New Directions for IR Evaluation: Online Conversations*, Sheffield, UK, July, pp. 9-15, 2004.

(Wong, 2004) Pak Chung Wong and Jim Thomas, “Visual Analytics,” *IEEE Computer Graphics and Applications*, 24(5)20-21, 2004.

(Yarowsky, 2003) David Yarowsky, “Scalable Elicitation of Training Data for Machine Translation,” *Team TIDES*, pp. 3-4, October, 2003.