

CROSS-LANGUAGE INFORMATION RETRIEVAL
ARIST CHAPTER
DOUGLAS W. OARD & ANNE R. DIEKEMA

INTRODUCTION

This chapter reviews research and practice in Cross-Language Information Retrieval (CLIR) that seeks to support the process of finding documents written in one natural language (e.g., English or Portuguese) with automated systems that can accept queries expressed in other languages. With the globalization of the economy and the continued internationalization of the Internet, CLIR is becoming an increasingly important capability that facilitates the effective exchange of information. For retrospective retrieval, CLIR allows users to state queries in their native language and then retrieve documents in any supported language. This can simplify searching by multilingual users and, if translation resources are limited, can allow monolingual searchers to allocate those resources to the most promising documents. In selective dissemination applications, CLIR allows monolingual users to specify a profile using words from one language and then use that profile to identify promising documents in many languages. Adaptive filtering systems that seek to learn profiles automatically can use CLIR to process training documents that may not be in the same language as the documents that later must be selected.

This review uses the term "documents" fairly broadly, since CLIR can be applied to a variety of modalities including character coded text, scanned images of printed pages, and recordings of human speech. Similarly, supporting the process of finding documents should be construed broadly as well, including both fully automated functions and capabilities that support productive human-system interaction. CLIR also appears in the literature as multilingual information retrieval (HULL & GREFENSTETTE), and as translanguing information retrieval (CARBONELL ET AL.), but all work conforming to the definition stated above is described in this chapter as CLIR for consistency.

The first reported work on CLIR was the development of the International Road Research Documentation system that used a controlled vocabulary thesaurus with aligned indexing terms in English, French and German (PIGUR). PEVZNER (1969, 1972) also implemented a Boolean exact match text retrieval system, translating a Russian thesaurus into English. SALTON (1970, 1973) conducted some smaller studies, augmenting the SMART system with hand-constructed bilingual term lists. By the mid-1970's it had been established that systems built using these techniques could achieve performance across languages on a par with their within-language performance. Commercial acceptance soon followed, and by 1977 ILJON was able to identify four multilingual text retrieval systems operating in Europe. Standardization quickly emerged as an important issue. In 1978 the International Standards Organization formally adopted ISO Standard 5964 on the construction of multilingual thesauri (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION), and that standard has remained unmodified since 1985.

Multilingual thesauri do not, however, completely solve the CLIR problem. DUBOIS identified three factors that motivate the search for other techniques: cost, currency and usability. First, indexing and maintenance costs limit the scalability of

thesaurus-based systems, although some automated tools are able to assist with these tasks. Second, thesauri in production applications often lag somewhat behind the current use of terminology because new words enter human languages each year. But perhaps the most serious limitation of thesaurus-based techniques is that untrained users seem to have difficulty exploiting their capabilities. Searching free text is the obvious alternative to use of a controlled vocabulary, and LANDAUER & LITTMAN (1990,1991) were the first to explore the potential for free text CLIR. Extending an automatic technique for reducing the effect of vocabulary differences on retrieval effectiveness, they sought to partially overcome the systematic vocabulary differences that result from choosing a different language. RADWAN & FLUHR began work in 1991 on an alternative technique that was based on translating the queries using manually encoded translation knowledge. Although much progress has been made since that time, these two early explorations of broad-coverage free text CLIR defined the two dominant themes that still guide research and practice: corpus-based and knowledge-based approaches.

Scope

This review brings together historical and contemporary research on automated techniques for cross-language retrieval of written and spoken text, both for retrospective retrieval and for selective dissemination. The review does not cover gestural languages such as American Sign Language, nor does it address language-independent techniques for recommending documents based either on ratings assigned by other users or on hypertext links. This is the first ARIST review of CLIR, but ERES previously reviewed international information transfer and METOYER-DURAN reviewed work on transfer of information across language barriers in a domestic context. Other surveys have addressed CLIR with more limited scope. FIGUR described early work on CLIR, with particular emphasis on developments in the former Soviet Union. FLUHR provided an overview of modern approaches, and OARD (1997b) provided a more recent overview. JONES & JAMES reviewed the field with particular attention to cross-language speech retrieval, and OARD & DORR (1996) produced the most extensive survey to date.

Organization

The review begins with an examination of the literature on user needs for CLIR. The main part of the chapter then follows the retrieval system model shown in Figure 1, adapted from OARD (1997c). Each section highlights the unique requirements imposed on one or more stages of that model in cross-language retrieval applications. The matching stage is covered in the somewhat greater detail, reflecting the treatment in the literature. Evaluation techniques are then described, and the review concludes with some observations regarding future research directions.

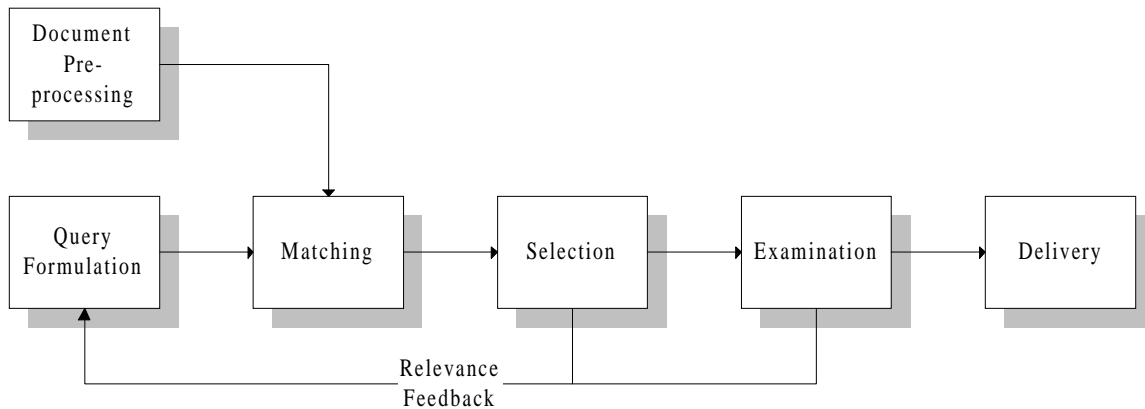


Figure 1. Information retrieval system model.

USER NEEDS

MEADOWS cited a number of studies that together suggest that in the early 1970's about half of the world's scientific literature was published in English. WELLISCH observed that English-language secondary sources (e.g., indexing and abstracting services) add to this total. Several massive translation efforts also contribute to the availability of information in English. World Translations Index, for example, lists 269 journals for which cover-to-cover translations (mostly into English) of every issue are prepared, and hundreds more that are selectively translated on a regular basis (INTERNATIONAL TRANSLATIONS CENTRE). STUDEMAN reported that the U.S. Foreign Broadcast Information Service translated 200 million words in a single year from over 3,500 publications in 55 languages. So English does serve, to some degree, as what WELLISCH called the "*lingua franca* of information retrieval tools."

Despite these efforts, much of the world's information is not available in English. HUTCHENS ET AL. found that about a third of the researchers at the University of Sheffield (United Kingdom) suspected that they had failed to learn of relevant work in a non-English speaking country. It turned out that a similar proportion had in fact discovered foreign language work that would have been more useful had it turned up earlier. MEADOWS found corroborating evidence for this problem on a larger scale, noting that researchers writing in English tend to over-cite other work in English and to under-use foreign language work, when compared to the linguistic distribution of scholarly writing in their field.

Discovering documents in a foreign language is, of course, only part of the problem. GOLDSTEIN found that between 20% and 45% of electrical engineers in Mexico encountered documents in unfamiliar languages at least once each month, and WOOD (1967) and ELLEN obtained similar results for a broad range of disciplines in the United Kingdom. WOOD (1974) offered some insight into the assistance that may be needed, reporting that over half the researchers requesting full-text translations from the British Library felt that summary translations of the results along with translations of figure and table captions were sufficient to provide the information that they required.

The recent growth of the global Internet has focused increased attention on the need for information exchange across linguistic barriers. A 1997 study by the INTERNET SOCIETY & ALIS TECHNOLOGIES, for example, found that 12% of World Wide Web pages that were randomly selected contained material in one of fourteen languages other than English. With upwards of 100 million web pages already indexed by the largest web search engines, this translates into an enormous potential demand for CLIR services. Projections during periods of exponential growth are always subject to question, but PIONEER CONSULTING estimates that by the year 2002 electronic collaborations will produce over 500 million messages per day that cross national borders.

DOCUMENT PREPROCESSING

Documents exist in many modalities, including character-coded text, printed pages and recorded speech. Each modality can, in turn, have several alternate representations. Character coded text can be expressed using different character sets, and a single character set may have alternative encodings. Some encoding schemes include alternate representations for the same character, and popular usage can introduce similar complications (e.g., accents on upper case Spanish characters are typically present in some countries but omitted in others). Similarly, printed pages may be available digitally in a number of formats, including page description languages or page images. For recorded speech, the speech rate can vary, a variety of accents may be present, and technical characteristics such as encoding and compression schemes can affect the fidelity of a recording. One goal of document preprocessing is to reduce this range of possible representations to a consistent character-coded text representation for each language that is present in a document.

Before such a representation can be constructed, the languages present in a document must be identified. This may be known *a priori*, it may be coded using a markup convention, or it may need to be determined from the contents of the document. GREFFENSTETTE compared two automatic language identification techniques for ten Western European languages. A technique based on the observed predominance of three-letter sequences (character trigrams) performed well, correctly classifying more than 93% of the evaluation sentences that contained at least six words and at least 99.8% of those that contained sixteen or more words. KIKUI integrated similar techniques with automatic character set detection for World Wide Web documents. ZISSMAN has shown that fairly accurate automatic spoken language identification is also possible, correctly classifying 89% of all 45 second speech samples as one of eleven languages. LEE ET AL. achieved correct language identification among six languages in 95% of scanned page images, a task complicated somewhat by the need for language-independent skew detection and by the variety of character fonts that might be used for each language.

Once the modality, language and encoding of each document are known, indexing features must be identified. In English, the most common indexing features for character coded text are word stems formed by automatic suffix removal. The utility of automatic suffix removal algorithms varies by language, with some languages exhibiting more productive morphology than English (and thus realizing a greater benefit from suffix removal) and other languages lacking any morphological variation at all. Construction of

a sophisticated stemmer for a new language might require considerable effort, but BUCKLEY ET AL. (1994) developed a useful Spanish stemmer in less than one day by manually identifying common suffix patterns. Short phrases that are detected by word cooccurrence, syntactic parsing, or dictionary lookup are sometimes indexed as well. Languages that permit fairly free construction of compositional compounds (e.g., German) make phrase detection straightforward, but compound splitting is then needed to identify constituent words within a compound term. For example, the German compound *Kraftwagenfuehrerschein* consists of *Kraftwagen* (truck or lorry) and *Fuehrerschein* (drivers license). WECHSLER ET AL. described a fairly effective compound splitting technique based on longest substring matching that requires only a list of terms for the language in question. Users might search for an entire phrase or compound, or they may search only for one of the constituent terms, so it is common to use both the phrase (or compound) and its constituent terms as indexing features. Some languages lack explicit word boundaries altogether in their written form (e.g., Chinese), introducing an extreme version of the compound splitting problem known as segmentation. GUO summarized prior work on Chinese segmentation and compared a number of techniques based on longest substring matching. WILKENSON has reported, however, that overlapping two character sequences (character bigrams) provided indexing features for Chinese that were about as effective as those discovered using dictionary-based longest substring matching.

Character recognition errors make the accurate identification of indexing features even more challenging when processing scanned page images. Optical character recognition systems typically depend on extensive training using manually assembled examples, and accurate systems are presently available for only a limited set of languages. Furthermore, recognition accuracy degrades rapidly when presented with poor reproductions or handwritten manuscripts. SMEATON & SPITZ sought to minimize these limitations while enhancing indexing speed by constructing indexing features for English words using encoded character shape groups (e.g., “b” and “h” might be assigned the same shape code) rather than individual character codes, and COOPER applied a similar technique to Thai. Retrieval effectiveness suffered significantly in English when compared to character coded text, but less of an adverse effect was apparent in Thai.

Recorded speech poses an even greater challenge, both with respect to speed and accuracy. Speech typically lacks explicit boundary markers between words, so the problem resembles the segmentation problem in languages such as Chinese, and variations in pronunciation, speaking rate, and the fidelity of the recording make speech recognition vastly more complex. Speech recognition systems trained on manually prepared time-synchronized transcripts can produce useful indexing features, but the needed training material is available for only a limited set of languages, recognition accuracy degrades when presented with applications for which the training material was not representative, and present processing speeds limit the size of the collections that can be indexed. SHERIDAN ET AL. (1997) used overlapping three-phoneme sequences (phone trigrams) as indexing features for recorded German speech in an attempt to overcome these limitations. Their initial results were disappointing, but NG & ZUE found have that phone trigrams can offer a viable alternative to word-based indexing for spoken documents in English.

QUERY FORMULATION

TAYLOR observed that users must compromise their information needs to match the perceived capabilities of available information systems when creating queries. Information retrieval systems seek to support this process by providing facilities for query specification and through incorporation of query refinement techniques such as relevance feedback. Users with little exposure to controlled vocabulary searching, for example, often find that formulation of effective queries using a printed thesaurus is difficult. Such users might benefit from a query interface that depicts the available indexing terms and their relationships in their preferred language. LI ET AL. developed such a system for English and Japanese, using versions of the INSPEC thesaurus in each language. Although no user study results were reported for multilingual applications, SMITH & POLLITT performed a qualitative assessment of a monolingual version of the same system.

The fully automatic query translation techniques described in the next section can be viewed as one type of support for query formulation in free text CLIR systems, but more interactive approaches have also been implemented. The QUILT system described by DAVIS & OGDEN (1997), for example, optionally displayed the Spanish translation of English query terms. A user who is able to read Spanish might thus be able to recognize erroneous translations, even if they lacked the fluency necessary to form effective queries without assistance. If so, they could then switch to a monolingual mode and enter the correct Spanish terms. YAMABANA ET AL. implemented a more sophisticated approach in which candidate translations of each term were displayed immediately, along with retranslations of each candidate back into the query language. Users unable to read the candidate translations could quickly skim the retranslations, and an alternate candidate could be chosen if necessary. An Internet demonstration of the READWARE system from Management Information Technologies, Inc. illustrated a further extension of this approach that could accommodate several languages simultaneously. READWARE depicted known senses of every query term using one near-synonym in the query language for each sense and allowed the user to designate the intended senses. For each selected word sense, a set of near-synonyms in English, German and French was passed on to the matching stage as a multilingual query. Together these three approaches illustrate a range of options that illustrate alternative ways of balancing capability with interface complexity.

MATCHING

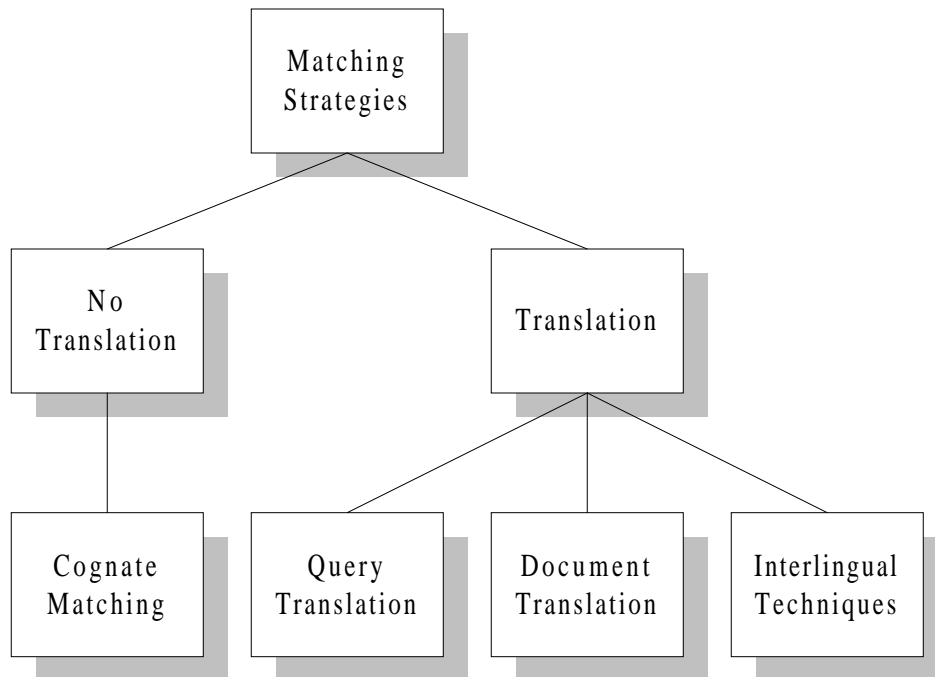


Figure 2. Matching strategies

Matching Strategies

Broadly stated, information retrieval systems construct representations of the documents and the information need and then match those representations to identify documents that are most likely to satisfy the need. In what MALONE ET AL. called "content-based" techniques, the representations are constructed from terms (e.g., stems, words, phrases, or character n-grams) that appear in the documents and the queries. Techniques for matching representations constructed from different vocabularies thus form a central component of CLIR systems. FURNAS ET AL. observed that information retrieval systems suffer from a vocabulary problem that results in part from variability in word usage. CLIR is simply an extreme case of this problem in which the words are selected from nearly disjoint vocabularies. Four general approaches to cross-language matching have emerged in CLIR: cognate matching, query translation, document translation, and interlingual techniques.

Cognate matching. Cognate matching essentially automates the process by which readers might try to guess the meaning of an unfamiliar term based on similarities in spelling or pronunciation. A simple version of cognate matching in which untranslatable terms are retained unchanged is often used in CLIR systems to match proper nouns and technical terminology (BALLESTEROS & CROFT 1997; GEY & CHEN, DAVIS & OGDEN, 1998; ELKATEB & FLUHR, HULL & GREFENSTETTE; KRAAIJ & HIEMSTRA). DAVIS extended this technique using fuzzy matching to discover Spanish cognates for English words that did not appear in a bilingual dictionary. BUCKLEY ET AL. (1998) applied a more sophisticated approach, creating equivalence classes for letter

sequences with similar sounds (e.g., “c,” “k,” and “qu” share an equivalence class). Since the translation knowledge is embedded directly in the matching scheme, cognate matching can be used in isolation. Most often, however, cognate matching is combined with other cross-language matching approaches.

Query Translation. Query translation is a more general strategy in which the query (or some internal representation of the query) is automatically converted into every supported language. Query translation is relatively efficient and can be done on the fly. The principal limitation of query translation is that queries are often short and short queries provide little context for disambiguation. Homonymous words (those with more than one distinct meaning) produce undesirable matches even in monolingual retrieval (KROVETZ & CROFT). Translation ambiguity compounds this problem, potentially introducing additional terms that are themselves homonymous. For this reason, controlling translation ambiguity is a central issue in the design of effective query translation techniques. Phrases typically exhibit less translation ambiguity than single words, and the literature suggests that phrase recognition strategies can substantially improve retrieval effectiveness. BALLESTEROS & CROFT (1997) observed beneficial effects from manual translation of phrases identified through syntactic analysis, and both RADWAN & FLUHR and KRAAIJ & HIEMSTRA explored techniques for automatically choosing an appropriate word order for phrases in which the constituent words had been translated separately. HULL & GREFFENSTETTE investigated the effect of noncompositional phrases that cannot be reconstructed from translations of the constituent terms and found an additional benefit.

Document translation. Document translation is just the opposite of query translation, automatically converting all of the documents (or their representations) into each supported query language. Documents typically provide more context than queries, so more effective strategies to limit the effect of translation ambiguity may be possible. Another potential advantage is that selected documents can be presented to the user for examination without on-demand translation (KRAAIJ). On the other hand, massive translation can be an expensive undertaking, and the costs are even greater if several query languages must be supported. As a result, relatively few experiments have compared document translation with query translation (OARD ET AL.), and ERBACH ET AL. suggested using document translation only for small collections in limited domains.

Interlingual techniques. Interlingual techniques convert both the query and the documents into a unified language-independent representation. Controlled vocabulary techniques based on multilingual thesauri are the most common examples of this approach. Because each controlled vocabulary term typically corresponds to exactly one concept, terms from any language may be used to index documents or to form queries. HLAVA ET AL. described a technique for partially automating the assignment of indexing terms to documents in several languages. Some fully automated interlingual techniques have also been implemented. Latent semantic indexing (LANDAUER & LITTMAN, 1990, 1991; DUMAIS ET AL.; BERRY & YOUNG; REHDER ET AL.) and the generalized vector space model (CARBONNELL ET AL.) both use a document aligned training corpus to learn a mapping from one or more languages into a language-neutral representation. Document and query representations from either language can be mapped into this space, allowing similarity measures to be computed both within and across languages.

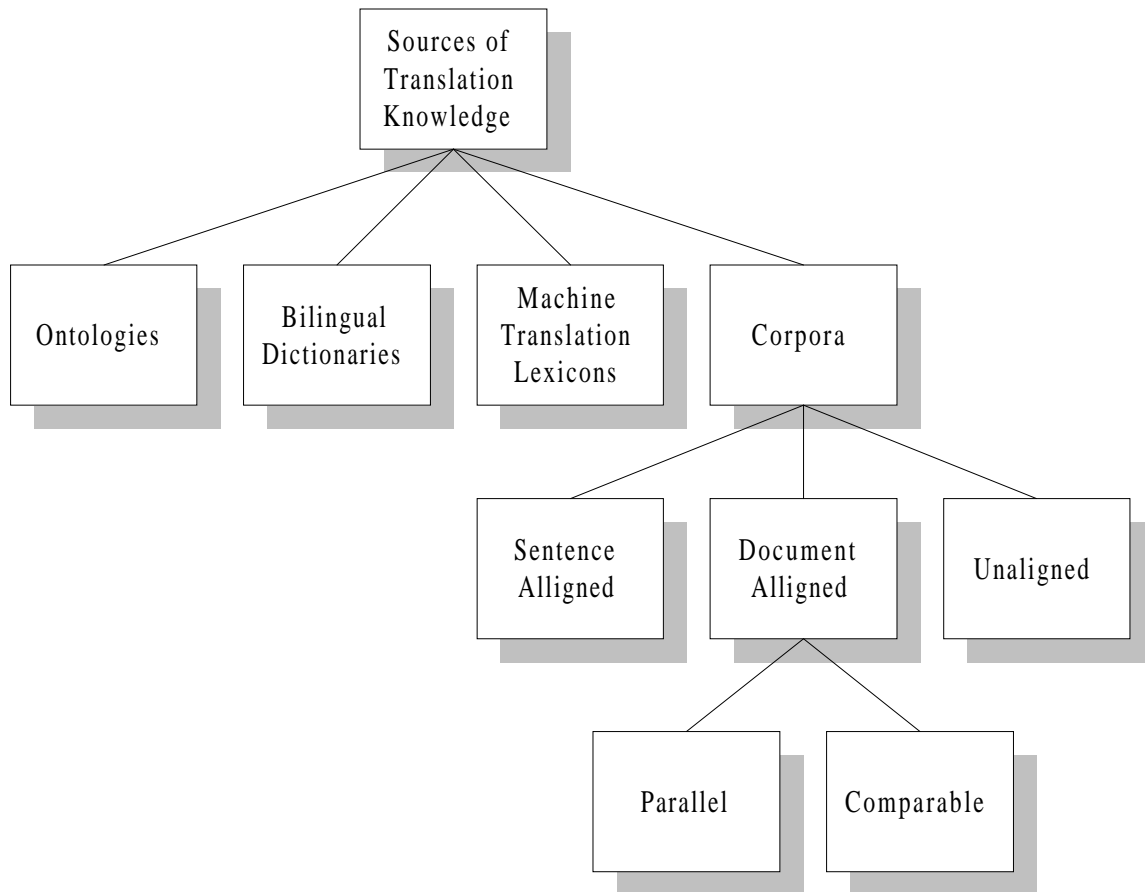


Figure 3. Sources of translation knowledge

Sources of Translation Knowledge

Each of the four matching approaches to CLIR depends on some form of translation knowledge. That knowledge may be encoded manually or extracted automatically from corpora, and CLIR techniques may take exploit translation knowledge in more than one form. The literature typically refers to techniques using translation knowledge from manually encoded translation knowledge as knowledge-based approaches. Techniques using translation knowledge from corpora are referred to as corpus-based techniques. The correspondence rules used for cognate matching represent one form of manually encoded translation knowledge. Three other manually encoded sources of translation knowledge have been applied to CLIR: ontologies, machine translation lexicons, and bilingual dictionaries. Three types of corpora have also been used: document-aligned corpora, sentence and term aligned corpora, and unaligned corpora. This section considers each of the six sources of translation knowledge in turn.

Ontologies. Ontologies are structures that encode domain knowledge by specifying relationships between concepts. Thesauri are ontologies that are designed specifically to support information retrieval. At present multilingual thesauri are the

dominant sources of translation knowledge in operational CLIR systems. Thesauri can support both controlled vocabulary and free-text retrieval, providing insight into both hierarchical relationships (broader terms, narrower terms), synonymy, and more general associations (related terms). Such relationships can help experienced users define better queries by enhancing their understanding of the structure of knowledge for the topic being searched. The European Parliament's multilingual EUROVOC thesaurus is one example of a multilingual thesaurus. A common approach to create a multilingual thesaurus is to translate an existing monolingual thesaurus, and KALACHKINA provides algorithms to deal with terms that lack direct translations. SOERGEL, however, cautions against merely translating an existing thesaurus since the expression of concepts in the original language will then dominate the conceptual structure. General-purpose ontologies such as WordNet (MILLER) are emerging as alternatives to traditional thesauri because their broader coverage permits use of sophisticated knowledge structures in broader domains that has heretofore been possible. By encoding additional relationships such as "part-whole" and "kind-of," WordNet explicitly captures a broader range of structural knowledge than traditional thesauri. The EuroWordNet project is developing a multilingual ontology resembling WordNet with components in Dutch, English, Italian and Spanish that are linked by an "interlingual index." CLIR support is a specific design goal of the project, and GILARRANZ ET AL. (1997a, 1997b) have described how EuroWordNet might be used to support a query translation strategy. Other projects (e.g., GermaNet described by HAMP & FELDWEG) are extending these ideas to other languages.

Bilingual dictionaries. Machine-readable bilingual dictionaries have been widely used to support query translation strategies (BALLESTEROS & CROFT, 1997; GEY & CHEN, DAVIS, DAVIS & OGDEN 1998; FLUHR ET AL., HULL & GREFENSTETTE; KRAAIJ & HIEMSTRA; KWOK; NGUYEN ET AL.; YAMABANA ET AL.). Bilingual dictionaries are typically designed for human use, so translations of individual terms are often augmented with examples showing how those terms could be used in context. It would be difficult to extract generalizations from those examples that could be used automatically, so machine readable dictionaries are typically processed manually or automatically to reduce them to a bilingual term list, perhaps with additional information such as part-of-speech. In essence, dictionary-based translation consists of looking up each query term in the resulting bilingual term list and selecting the appropriate translation equivalents. The simplest way of using such a bilingual term list is to select every known translation for each term, and that approach is often used as a baseline in dictionary-based CLIR evaluations. Both RADWAN & FLUHR and DAVIS have shown that limiting the translations to those with the same part-of-speech (e.g., noun or verb) can improve retrieval effectiveness, and KRAAIJ & HIEMSTRA experimented with the use of preferred translations that were noted in their dictionary. OARD ET AL. demonstrated that arbitrarily choosing a single translation can be just as good (by the average precision measure), apparently because on balance as many queries are helped as are hurt. HULL explored the ability of structured queries to further limit translation ambiguity, implementing a weighted Boolean matching strategy that exploited the observation that correct translations are more likely to cooccur than incorrect translations. Dictionary-based CLIR can suffer from limited dictionary coverage, inaccuracies during automatic

construction of the bilingual term list, and incorrect selection of the appropriate translation equivalents (BALLESTEROS & CROFT, 1997; FLUHR ET AL.; GAUSSIET ET AL.; HULL & GREFENSTETTE; NGUYEN ET AL.), but it is sufficiently efficient and effective to be useful in many applications.

Machine translation lexicons. Machine translation systems are becoming fairly widely available, although machine-readable dictionaries still cover a greater number of language pairs (KRAAIJ). Machine translation systems encode translation knowledge in a “lexicon” that contains the information needed for automatic analysis, translation and generation of natural language. One goal of natural language analysis is to disambiguate terms in ways that can limit translation ambiguity, and the lexicon is often designed to provide information that is useful for this purpose. The most straightforward way to apply a machine translation lexicon to CLIR is to simply use the machine translation system to translate either the queries or the document collection. Queries are rarely provided as well formed sentences, however, so the effectiveness of this approach may be limited in query translation applications (HULL & GREFENSTETTE, KRAAIJ). Machine translation systems necessarily choose a single preferred translation for each term, and ERBACH ET AL. have observed that such a singular choice might adversely affect retrieval effectiveness. Examples of the use of machine translation for query and document translation can be found in OARD & HACKETT.

Document aligned corpora. Document aligned corpora are document collections in which useful relationships between sets of documents in different languages are known. Parallel corpora are made up of translation equivalent sets, each containing a document and one or more translations. Comparable collections, on the other hand, are typically separately authored but related by topical content. Aligned document sets in comparable corpora may contain one or more documents in each language (PETERS & PICCHI; SHERIDAN & BALLERINI). The basic strategy for using document aligned corpora is to represent each term using the pattern of aligned sets in which that term occurs and then to construct language-neutral representations of documents in any supported language using the resulting term representations. Techniques from linear algebra are typically used to compute and manipulate these term representations. When the language of each document is known, each term is typically tagged with a language marker in order to avoid undesired conflation of different concepts in other languages. CARBONELL ET AL. implemented one such technique, the Generalized Vector Space Model (GVSM), using a parallel corpus. Latent Semantic Indexing (LSI) extends this approach by conflating terms that have similar representations, often increasing recall without adversely affecting precision. Both parallel corpora (LANDAUER & LITTMAN 1990, 1991; DUMAIS ET AL.) and comparable corpora (REHDER) have been used with LSI. BARRY & YOUNG found that the effectiveness of LSI could be improved by using an aligned corpus of short passages rather than one formed from longer documents. Although LSI is sometimes more effective than GVSM, computation of the term conflation step is computationally intensive (CARBONELL ET AL.). SHERIDAN & BALLERINI and MATEEV ET AL. investigated an alternative approach, building a bilingual term list for query translation using term representations computed from a comparable corpus of news stories that was aligned using classification codes, publication dates and cognates. They found the terms in each language that were most similar to each query term (using a

vector similarity measure) and then used several of the most similar terms as the translated query. While LSI uses more sophisticated techniques to conflate similar terms, SHERIDAN & BALERINI's technique is more efficient.

Sentence and term aligned corpora. Comparable corpora can be aligned only to the document level, but many individual sentences in parallel corpora can be aligned automatically using dynamic programming techniques. DAVIS used a sentence-aligned parallel corpus directly to augment dictionary based query translation without substantial improvement over a simpler dictionary-based technique. OARD (1996, 1997a) used sentence alignments as a basis for aligning individual terms, but again found that knowledge based techniques (in this case, machine translation) were more effective when the corpus based technique was required to extract translation knowledge from one collection and then apply it to another. In those experiments, a set of sentence aligned translations of United Nations documents was used as a source of translation knowledge, and a monolingual collection of Spanish newswire articles was used for evaluation. CARBONELL ET AL. implemented a similar approach, evaluating retrieval effectiveness on a portion of the same corpus from which translation knowledge had been extracted. Under those conditions, the sentence aligned corpus that was used to produce term alignments outperformed every other technique they tried. OARD (1997a) saw a similar improvement when comparing the same-collection performance of LSI with the performance of the same algorithm when trained on a different collection. It thus appears that document and sentence aligned techniques may be most useful when the needed alignments are known within some portion of the same collection from which retrieval is desired. Although such a situation may exist in a few applications (e.g., if translations are being made routinely, but they are not available immediately), this factor is likely to somewhat circumscribe the utility of techniques based on document and sentence aligned corpora.

Unaligned corpora. A representative monolingual document collection is, of course, available in any in application of CLIR to retrospective retrieval. Such collections are often assembled for filtering applications as well because they provide useful collection frequency statistics. When representative documents in more than one language are present in (or can be added to) such a collection, the collection itself can be used in conjunction with a bilingual term list as an additional source of translation knowledge even if *a priori* document alignments are not known. BALLESTEROS & CROFT (1997) applied fully automatic passage-level pseudo-relevance feedback using the query language portion of their unaligned corpus to refine the query representation. By augmenting the original query with terms appearing in top-ranked passages, monolingual pseudo-relevance feedback often improves recall without a significant adverse effect on precision. They then applied dictionary based query translation to produce a version of the query in the desired language, followed by fully automatic passage-level pseudo-relevance feedback using the portion of the unaligned corpus containing documents in that language. When applied individually, each pseudo-relevance feedback step improved CLIR effectiveness, and the combination outperformed either step alone. KROVETZ & CROFT and SANDERSON have shown that ranked retrieval techniques tend to reinforce the appropriate interpretation of words that admit more than one interpretation. Viewed in this light, the first pseudo-relevance feedback step serves to limit the adverse effect of

translation ambiguity by including additional terms that are related to the original query terms. YAMABANA ET AL. sought to achieve the same result more directly. For each query term, they identified one related term in the unaligned corpus that often appeared in a sentence with the query term. They then selected the candidate that most often appeared in the same sentence as some possible translation of the related term. YAMABANA ET AL. obtained some improvement in translation accuracy using this technique, but they did not evaluate the effect of that improvement on retrieval effectiveness. PICCHI & PETERS have proposed a similar technique that exploits more context by considering the possible translations of groups of words surrounding each query term in the unaligned corpus. Although techniques based on unaligned corpora appear promising, SHERIDAN ET AL. (1997) failed to find any improvement when using languages and collections different from those used by BALLESTEROS & CROFT. It thus appears the nature of the unaligned corpus and/or the way in which additional context-revealing terms context are chosen can substantially affect the results.

SELECTION, EXAMINATION AND DELIVERY

As MARCHIONINI has observed, searching and browsing are complementary activities. Automated systems apply rather simple techniques to enormous volumes of information, while humans can effectively exploit quite sophisticated selection heuristics on fairly small sets. One important goal of the user interface is to expose the information on which users can base these decisions. Retrieval systems containing full text typically support two browsing strategies: selection of documents from a list of promising candidates identified by the system, and detailed examination of individual documents.

Support for selection presents unique challenges when the documents are written in an unfamiliar language. Monolingual selection interfaces typically present document titles along with some information about the source of the material and when it was produced. Occasionally some form of summary such the first few lines or some individual words automatically extracted from the document are also presented. Conversion of names and dates using simple transliteration schemes is relatively straightforward, but title translation is more complex. Translation of titles using a fully automatic machine translation is a possibility, but titles rarely form the sort of well-formed linguistic expressions that typical machine translation systems are optimized for. KIKUI ET AL. reported that choosing the most common candidate translation (using a monolingual corpus) and then reordering the terms using some simple rules produced usable translations of English web page titles into Japanese. RESNIK evaluated an alternative strategy for translating brief listings into English, displaying as many as three alternative translations when faced with translation ambiguity. Using a decision theoretic measure, they found that such translations were more effective than a naive Bayesian classifier, but not as effective as monolingual selection.

Support for examination poses an even greater challenge. Several companies market translation software that is compatible with popular web browsers, and proxy translation servers are becoming available on the Internet. Typical machine translation systems are not yet fast enough to keep up with interactive selection and scrolling behavior, however, so interactive searching is inhibited to some extent when query

translation is used. Approaches based on advance translation of every document avoid this problem, but the time and expense involved limit application of those techniques. Rapid word-by-word translation like that explored by KIKUI ET AL. and RESNIK could in principle be used with query translation, but the utility of such techniques for examining relatively long documents in a CLIR system has not yet been explored. Traditional abstracting services such as INSPEC have adopted a more parsimonious approach, manually preparing abstracts for every document in the supported query language (usually English) regardless of the abstracted documents' language. FRANZEN & KARLGREN (1997) proposed automating this process by translating brief extracts or summaries as an alternative to translating entire documents on demand, but research on cross-language summarization is just beginning.

The ultimate delivery of selected documents in a usable form may be a somewhat more tractable problem than support for interactive examination if adequate time for translation can be allowed when arranging for delivery. O'HAGAN provided an overview of the translation industry and observed that globally interconnected networks will make it possible to marshal worldwide translation resources upon demand. Although fully automatic machine translation can presently only produce high quality translations in very limited subject areas, O'HAGAN suggested that a robust and responsive translation infrastructure could be built using machine assisted human translation. The human effort involved will likely make delivery the most expensive component on a per-document basis, so effective recognition of the most promising documents using the query formulation, matching, selection and examination stages is particularly important.

EVALUATION

Experimental evaluation of CLIR systems poses unique challenges because the languages covered by the translation resources must match the languages covered by the evaluation resources. The situation is further complicated when alternative techniques that require different translation resources are compared. A CLIR test collection thus consists of a set of documents in one or more languages, a set of queries in a language or languages different from that of the documents, relevance judgments for each query-document pair, and translation resources such as dictionaries, bilingual corpora, or cognate matching rules.

LANDAUER & LITTMAN (1990) developed a simple evaluation technique known as mate finding for use with document-aligned corpora. Mate finding is a variation on known item retrieval, a classic evaluation strategy in which the rank assigned to a unique item that is known to be relevant to the query is used as the measure of effectiveness. LANDAUER & LITTMAN (1990, 1991) partitioned an English-French parallel collection, extracting translation knowledge from one part and using the other part for evaluation. Each English document was then used as a query, and statistics describing the rank of the known French translation for each document were presented. CARBONELL ET AL. found that mate retrieval was less able to discriminate among fairly good techniques than more traditional strategies in which recall and precision were reported, but mate retrieval remains useful as a simple strategy for identifying promising CLIR techniques when more sophisticated evaluation resources are not available.

RADWAN & FLUHR used French translations of the 1,398 abstracts in the English Cranfield collection to compute precision-recall graphs and an average precision measure. DAVIS & DUNNING adopted an alternate strategy, manually translating Spanish topic descriptions into English and then using those topic descriptions to construct English queries to retrieve Spanish newswire articles from the Text REtrieval Conference (TREC). Manual translation of queries is now a widely used evaluation strategy because it permits existing test collections to be inexpensively extended to any language pair for which translation resources are available. Because manual translation requires the application of human judgment, evaluation collections constructed in this way exhibit some variability based on the terminology chosen by a particular translator. But if a standard set of translations is agreed upon, such a strategy offers a meaningful basis for selecting between alternative CLIR techniques.

There are, however, some applications for which manual query translation would not produce an adequate test collection. Corpus-based techniques, for example, may not perform well on collections that differ markedly from the corpora on which they were trained. There presently is no widely accepted metric for reporting the similarity of two corpora, so same-corpus (i.e., best case) evaluations are typically performed using a held-back portion of the corpus. CARBONELL ET AL. produced a test collection in this way by exhaustively performing 33,630 relevance judgments for a portion of a parallel collection of English and Spanish documents. This produced a test collection that was about the same size as the Cranfield collection used by RADWAN & FLUHR, but with the added characteristic that the remainder of the parallel corpus was available for the extraction of translation knowledge. SHERIDAN & BALLERINI also built a test collection from a document aligned corpus, but they developed a genre-specific strategy for newswire articles. By constructing queries for unpredicted events and ending their search three days after the event (which produced a different collection size for each query) they cut down the number of relevance judgments considerably. Newswire stories are fairly readily available in character-coded form, so this evaluation strategy may provide an economical alternative for many applications.

Evaluation of adaptive filtering techniques that learn to select documents in one language based on user reactions to documents in other languages imposes further requirements on an evaluation collection because a third partition of the evaluation collection may be needed. OARD (1997a) constructed such a collection using monolingual test collections in English and Spanish for which four topic descriptions were closely aligned and a parallel corpus of English and Spanish documents for which no relevance judgments were needed. In addition to the adaptive filtering evaluation, some indication of the degree of similarity between one of the monolingual test collections and the parallel corpus was also obtained.

Relatively large document collections are needed to accurately reflect the performance of IR systems in large-scale applications, and potential need to subdivide the collection two or three ways exacerbates the situation. Obtaining statistical significance will often require more queries for query translation experiments than for monolingual experiments on the same collection because uneven translation accuracy introduces an additional source of variation. And collections covering a wide range of languages and modalities will be needed to assess the effect of variations in morphology, word boundary

marking, and recognition accuracy. At present the TREC CLIR collection described by MATEEV ET AL. and SCHÄUBLE & SHERIDAN is the most comprehensive step in that direction. Using an approach known as pooled relevance assessment, relevance judgments for about 100,000 newswire articles in each of three languages (English, French and German) were developed by judging documents selected using several different retrieval techniques. The documents are not translations of each other, but they are drawn from the same genre and time frame and SHERIDAN ET AL. (1998) have automatically identified some possible alignments between some of the French and German documents in the collection.

Some insight into the contribution of alternative translation techniques can be obtained by comparing CLIR results with the effectiveness of a similar monolingual technique on the same collection. Typically expressed as a percentage of monolingual effectiveness, reported values typically range from around 50% for unconstrained dictionary based query translation to 75% or so for more sophisticated techniques. Direct comparisons are difficult, however, because the monolingual reference technique is often different, parameter variations can introduce additional variations even when the reference technique is nominally the same, the effect of differing collections on relative effectiveness is not well characterized, and different effectiveness measures may have been used. HULL & GREFENSTETTE reported precision averaged over several fixed numbers of documents to characterize high precision interactive searching, while BALLESTEROS & CROFT (1997) reported precision averaged over the full range of recall values. Relative performance figures can help identify particularly promising techniques, however, and then the most promising techniques can be subjected to a more rigorous side by side comparison.

RESEARCH DIRECTIONS

Nearly three decades of research on and practice of controlled vocabulary techniques for CLIR and eight years of research on free text techniques have produced a wide array of useful techniques, but more remains to be done. Existing research on user needs for CLIR, for example, addresses the deliberate dissemination of information well but the impact of ubiquitous networking and the resulting trend towards flattened organizational structures has yet to be addressed. Some issues, such as the impact of networked communications on the translation infrastructure supporting ultimate use of selected documents, have implications for both controlled vocabulary and free text CLIR. But free text techniques are still relatively new, and it is there that many of the open research questions are to be found.

Important research issues are found in each stage of the model shown in Figure 1. The distinction between user-assisted and fully automatic query translation is rather sharply drawn at present, with users either being offered the opportunity to help resolve translation ambiguity for every term or for none of them. More sophisticated strategies might retain much of the benefit of user-assisted translation while avoiding unnecessary allocation of user effort and screen space to that task. Present document preprocessing systems are typically language specific, often using hand-built components for tasks such as character set conversion, compound splitting, and stemming. The development of

easily configured tools for such tasks would make the addition of additional languages a far more tractable task. The matching stage has received a great deal of attention, but cognate matching has only recently been investigated carefully. Further work on additional language pairs and strategies for combining cognate matching with other techniques appear to be the natural next steps. The importance of selection, examination and delivery for CLIR system design is now beginning to be recognized, but much remains to be done. It is not yet clear, for example, whether rapid translation of the entire text or automatic generation of translated summaries will provide the best support for examination, and answering that question may require the development of new evaluation techniques. Other evaluation issues also require attention. Perhaps most importantly, it will not be possible to accurately characterize the performance of document and sentence aligned corpus-based techniques in practical applications without some way to measure the degree of difference between the corpus from which the translation knowledge is extracted and the collection from which retrieval is desired.

As CLIR has matured, increasingly integrated approaches have been investigated. Dictionary based query translation has been improved using unaligned corpora (BALLESTEROS & CROFT 1996,1997), and term aligned corpora have been refined by seeding the alignments using a bilingual dictionary (YANG ET AL. 1997). Fully automatic query translation techniques are being augmented with user assisted query translation. This trend will likely continue, encompassing other components and techniques as productive interactions are discovered. GACHOT ET AL., for example, has observed that closer coupling between machine translation and matching techniques might be helpful because additional linguistic information would be available. Ultimately the distinctions that have been drawn in this chapter between separate components and different techniques may be as useful for explaining how they are coupled as for how they are different.

CONCLUSION

Controlled vocabulary CLIR techniques are now widely deployed, and free text systems for practical applications are beginning to appear. Although monolingual retrieval is still more effective for free text than CLIR, several useful CLIR techniques are known. Query translation, document translation, interlingual techniques and cognate matching provide a range of alternatives that can be tailored to specific applications. Document preprocessing strategies have been developed for scanned page images and recorded speech, but character coded text remains the most easily processed format. Interactive applications pose additional challenges, since users may not have the language skills that would be needed to select and examine documents in their original language. Additional opportunities are present as well, however, since the user can help refine translation knowledge that is extracted from dictionaries, bilingual corpora, or other sources. Evaluation poses additional challenges that the recent development of the TREC CLIR test collection has begun to address.

Many modern information systems support only a single language, but that limitation will likely become increasingly untenable in an era of ubiquitous global networks and vast international information flows. Cross-language information retrieval is one

component of the technological infrastructure that will help make the World Wide Web a truly worldwide resource, and it will undoubtedly find widespread application in other parts of the information industry as well. Although much remains to be done, the techniques that have been developed and the ways in which they have been applied provide useful signposts for developers that wish to begin exploring the opportunities that cross-language information retrieval presents.

BIBLIOGRAPHY

[There is still some formatting work to do, and some details such as ISSN, ISBN, and page numbers are missing, but everything is here.]

ALLAN, J., CALLAN, J., CROFT, W. B., BALLESTEROS, L., BYRD, D., SWAN, R., & XU, J. 1998. INQUERY Does Battle with TREC-6. In: Proceedings of the Sixth Text Retrieval Conference (TREC-6), National Institute of Standards and Technology (NIST), Gaithersburg, MD. <http://trec.nist.gov>

ATA, B. M. A., MOHD, T., SEMBOK, T., & YUSOFF, M. 1995. SISDOM : a multilingual document retrieval system. *Asian Libraries*, 1995; 4(3): 37--46

AUSTIN, D. 1977. Progress Towards Standard Guidelines for the Construction of Multilingual Thesauri. In: *Third European Congress on Information Systems and Networks (Vol. 1)*: Verlag Dokumentation. 1977. pp. 341--402.

BALLESTEROS, L., & CROFT, W. B. 1996. Dictionary Methods for Cross-Lingual Information Retrieval. In: R. R. Wagner & H. Thoma (Eds.) *New York: Springer*. 1996. pp. 791--801. ISBN: 354061656X. Also appeared in *Lecture Notes in Computer Science*, ISSN: 0302-9743 1996, issue 1134. <http://ciir.cs.umass.edu/info/psfiles/irpubs/ir.html>

BALLESTEROS, L., & CROFT, W. B. 1997. Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In: *Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval* .1997. pp. ?

BENKING, H., & KAMPPFMEYER, U. 1992. Harmonization of Environmental Meta-Information with a Thesaurus-based multi-lingual and multi-medial Information System. In: A. Zygielbaum (Ed.), *AIP Conference Proceedings 283, Earth and Space Science Information Systems*: American Institute of Physics. 1992. pp. 688--695.

BERRY, M., & YOUNG, P. 1995. Using Latent Semantic Indexing for Multilanguage Information Retrieval. *Computers and the Humanities*, 1995; 29(6): 413-429. ISSN 0010-4817.

BLAKE, P. 1992. The MenUSE System for Multilingual Assisted Access to Online Databases, in the context of current EC funded projects. *On-line Review*, 1992; 16(3): 139-146. June. ISSN 0309-314X.

BUCKLEY, C., MITRA, M., WALZ, J., & CARDIE, C. 1998. Using Clustering and SuperConcepts Within SMART: TREC 6. In: *Proceedings of the Sixth Text Retrieval Conference (TREC-6)*, National Institute of Standards and Technology (NIST), Gaithersburg, MD. <http://trec.nist.gov>

BUCKLEY, C., SALTON, G., ALLAN, J., & SINGHAL, A. 1994. Automatic Query Expansion Using SMART: TREC 3. In: Harman, D. K. (Ed.), Overview of the Third TextREtrieval Conference (TREC-3), pp. 69-80. National Institute of Standards and Technology (NIST), Gaithersburg, MD.
<http://www-nlpir.nist.gov/TREC/trec3.papers/cornall.new.ps>

CARBONELL, J., YANG, Y., FREDERKING, R., BROWN, R. D., GENG, Y., & LEE, D. 1997. Translingual Information Retrieval: A Comparative Evaluation. In: Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence. pp. ??

CHACHRA, V. 1993. Subject Access in an Automated Multithesaurus and Multilingual Environment. In: S. McCallum & M. Ertel (Eds.), 2nd Satellite Meeting on Automated Systems for Access to Multilingual and Multiscript Library Materials : Saur.1993. pp. 63--76. ISBN: 3598217978; Also appeared in IFLA PUBLICATIONS 1994, Vol. 70.

CHMIELEWSKA-GORCZYCA, E., & STRUK, W. 1994. Translating Multilingual Thesauri. In: P. Stanucikova & I. Dahlberg (Eds.), 1st European Conference on Environmental Knowledge Organization and Information Management. Frankfurt: Indeks Verlag.1994. pp. 150--155. ISBN: 3886726002 3886726010; also appeared in Knowledge Organization in Subject Areas ISSN 0946-9389, 1994, vol. 1.

COOPER, D. 1997. How to Read Less and Know More: Approximate OCR for Thai. In: Belkin, N., Narasimhalu, D. & Willett, P. (Eds.), Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval: ACM SIGIR.1997. pp. 216-225. ISBN 0-89791-836-3.

D'OLIER, J. H. 1977. Multilingualism in Scientific and Technical Documentation. International Forum on Information and Documentation, 1977; 2(4): 20--24

DAVIS, M. 1997. New Experiments in Cross-Language Text Retrieval at NMSU 's Computing Research Lab. In: D. K. Harman (Ed.), The Fifth Text REtrieval Conference (TREC-5). National Institute of Standards and Technology (NIST), Gaithersburg, MD.
<http://crl.nmsu.edu/users/madavis/Site/Book2/trec5.ps>

DAVIS, M. & DUNNING, T. 1995. A TREC Evaluation of Query Translation Methods for Multi-Lingual Text Retrieval. In: Harman, D. K. (Ed.) The Fourth Text Retrieval Conference (TREC-4). National Institute of Standards and Technology (NIST), Gaithersburg, MD.
<http://trec.nist.gov>, <http://crl.nmsu.edu/users/madavis/Site/Book2/trec4.ps>

DAVIS, M. W., & OGDEN, W. C. 1997. Implementing cross-language text retrieval systems for large-scale text collections and the world wide web, AAAI Symposium on Cross-Language Text and Speech Retrieval : American Association for Artificial Intelligence. 1997. pp. 2-10. ISBN: 1-57735-040-5; Technical Report: SS-97-05.
<http://www.clis.umd.edu/dlrg/filter/sss/papers/>

DAVIS, M. & OGDEN, W. 1998. Free Resources and Advanced Alignment for Cross-Language Text Retrieval. In: Proceedings of the Sixth Text Retrieval Conference (TREC-6), National Institute of Standards and Technology (NIST), Gaithersburg, MD.
<http://trec.nist.gov>

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY 1996. Tipster Text Program. Morgan Kaufmann

DUBOIS, C. P. R. 1987. Free Text vs. Controlled Vocabulary: A Reassessment. Online Review, Vol. 11, No. 4, pp. 243-253.

DUCLOY, J. 1996. Tools and Techniques for Digital Libraries. ERCIM News, 1996; 27.
http://www-ercim.inria.fr/publication/Ercim_News/enw27/ducloy.html

DUMAIS, S. T., LETSCHE, T. A., LITTMAN, M. L., & LANDAUER, T. K. 1997. Automatic Cross-Language Retrieval Using Latent Semantic Indexing, AAAI Symposium on Cross-Language Text and Speech Retrieval : American Association for Artificial Intelligence. 1997. pp. 15-21. ISBN: 1-57735-040-5; Technical Report: SS-97-05.
<http://www.clis.umd.edu/dlrg/filter/sss/papers/>

ELKATEB, F. & FLUHR, C. 1998. EMIR at the CLIR Track of TREC 6. In: Proceedings of the Sixth Text Retrieval Conference (TREC-6), National Institute of Standards and Technology (NIST), Gaithersburg, MD. <http://trec.nist.gov>

ELLEN, SANDRA R. 1979. Survey of Foreign Language Problems Facing the Research Worker. Interlending Review, Vol. 7, No. 2, pp. 31-41, April.

ERES, B. K. 1989. International Information Issues. In: Williams, M. (Ed.) Annual Review of Information Science and Technology, Vol. 24, p. 3

ERBACH, G., NEUMANN, G., & USZKOREIT, H. 1997. MULINEX Multilingual Indexing Navigation and Editing Extensions for the World-Wide Web, AAAI Symposium on Cross-Language Text and Speech: American Association for Artificial Intelligence. 1997. pp. 22-28. ISBN: 1-57735-040-5; Technical Report: SS-97-05.
<http://www.clis.umd.edu/dlrg/filter/sss/papers/>

EVANS, D. A., HANDERSON, S. K., MONARCH, I. A., PEREIRO, J., DELON, L., & HERSH, W. R. 1991. Mapping Vocabularies Using "Latent Semantics" (CMU-LCL-91-1): Carnegie Mellon University, Laboratory for Computational Linguistics

FLUHR, C., & RADWAN, K. 1993. Fulltext Databases as Lexical Semantic Knowledge for Multilingual Interrogation and Machine Translation. In: P. Brezillon & V. Stefanuk (Eds.), Proceedings of the East-West Conference on Artificial Intelligence (EWAIC '93) Moscow: Association for Artificial Intelligence of Russia, ICSTI.1993. pp. 124--128.

FLUHR, C. 1995. Multilingual Information Retrieval. In: R. A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, & V. Zue (Eds.), Survey of the State of the Art in Human Language Technology : Center for Spoken Language Understanding, Oregon Graduate Institute.1995. pp. 391--305. <http://www.cse.ogi.edu/CSLU/HLTsurvey/ch8node7.html>

FLUHR, C., SCHMIT, D., ELKATEB, F., ORTET, P., & GURTNER, K. 1997. Multilingual Database and Crosslingual Interrogation in a Real Internet Application, AAI Symposium on Cross-Language Text and Speech Retrieval : American Association for Artificial Intelligence.1997. pp. 32-36. ISBN: 1-57735-040-5; Technical Report: SS-97-05. <http://www.clis.umd.edu/dlrg/filter/sss/papers/>

FRANZEN, K., & KARLGREN, J. 1997. Project Presentation REPTILE Retrieval Extraction Presentation and Translation using Language Engineering, AAI Symposium on Cross-Language Text and Speech Retrieval : American Association for Artificial Intelligence.1997. pp. 37-39. ISBN: 1-57735-040-5; Technical Report: SS-97-05. <http://www.clis.umd.edu/dlrg/filter/sss/papers/>

FREDERKING, R., MITAMURA, T., NYBERG, E., & CARBONELL, J. 1997. Translingual Information Access, AAI Symposium on Cross-Language Text and Speech Retrieval : American Association for Artificial Intelligence.1997. pp. 40-48. ISBN: 1-57735-040-5; Technical Report: SS-97-05. <http://www.clis.umd.edu/dlrg/filter/sss/papers/>

FURNAS, G. W., LANDAUER, T. K., AND GOMEZ, L. M., & DUMAIS, S. 1987. The Vocabulary Problem in Human-system Communication. Communications of the Association for Computing Machinery. Vol. 30, No. 11, pp. 964-971.

GACHOT, D. A., LANGE, E., & YANG, J. 1998. The SYSTRAN NLP Browser: An Application of Machine Translation Technology in Multilingual Information Retrieval. In: G. Grefenstette, (Ed.), Cross Language Information Retrieval: Kluwer Academic. pp. ?. ISBN:0-7923-8122-X, pp. ?. <http://www.rxc.xerox.com/research/mltt/DMHead/CLIR/>

GAUSSIÉ, E., GREFENSTETTE, G., HULL, D. A., & SCHULZE, B. M.1998. Xerox TREC-6 Site Report: Cross Language Text Retrieval. In: Proceedings of the Sixth Text Retrieval Conference (TREC-6), National Institute of Standards and Technology (NIST), Gaithersburg, MD. <http://trec.nist.gov>

GEY, F. & CHEN, A. 1998. Phrase Discovery for Cross-Language Retrieval at TREC 6. In: Proceedings of the Sixth Text Retrieval Conference (TREC-6), National Institute of

Standards and Technology (NIST), Gaithersburg, MD. <http://trec.nist.gov>

GIBB, J. M., & PHILLIPS, E. 1977. Scientific and Technical Publishing in a Multilingual Society. In: Third European Congress on Information Systems and Networks 1977. pp. 13--27.

GILARRANZ, J., GONZALO, J., & VERDEJO, F. 1997a. An approach to Conceptual Text Retrieval Using the EuroWordNet Multilingual Semantic Database, AAAI Symposium on Cross-Language Text and Speech Retrieval: American Association for Artificial Intelligence.1997. pp. 49-55. ISBN: 1-57735-040-5; Technical Report: SS-97-05. <http://www.clis.umd.edu/dlrg/filter/sss/papers/>

GILARRANZ, J., GONZALO, J., & VERDEJO, F. 1997b. Language-Independent Text Retrieval with the EuroWordNet Multilingual Semantic Database. In: Second Workshop on Multilinguality in the Software Industry: The AI Contribution.1997. pp. ? <http://www.iit.nrcps.ariadne-t.gr/~costass/mulsaic97.html>

GOLDSTEIN, E. S. 1985. The Use of Technical Information by Engineers of the Electrical Sector of Mexico. Unpublished Doctoral Dissertation, University of California, Los Angeles.

GREFENSTETTE, G. 1995. Comparing Two Language Identification Schemes. In. Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data. <http://www.rxrc.xerox.com/researc/mltt/Tools/guesser.html>

GREFENSTETTE, G. 1998. Cross Language Information Retrieval: Kluwer Academic. ISBN:0-7923-8122-X

GUO, J. 1997. A Comparative Study on Sentence Tokenization Generation Schemes. In review for journal publishing, January, 1997. <http://sunzi.iss.nus.sg:1996/guojin/papers/>

HAMP, B. & FELDWEG, H. GermaNet – A Lexical-Semantic Net for German. <http://www.sfs.nphil.uni-tuebingen.de/isd/english.html>

HAYASHI, Y., KIKUI, G. I., & SUSAKI, S. 1997. TITAN: A Cross-Linguistic Search Engine for the WWW, AAAI Symposium on Cross-Language Text and Speech Retrieval: American Association for Artificial Intelligence.1997. pp. 56-62. ISBN: 1-57735-040-5; Technical Report: SS-97-05. <http://www.clis.umd.edu/dlrg/filter/sss/papers/>

HLAVA, M. M. K., HAINEBACH, R., BELONOGOV, G., & KUZNETSOV, B. 1997. Cross-Language Retrieval - English/ Russian/ French, AAAI Symposium on Cross-Language Text and Speech Retrieval: American Association for Artificial Intelligence.1997. pp. 63-83. ISBN: 1-57735-040-5; Technical Report: SS-97-05.

<http://www.clis.umd.edu/dlrg/filter/sss/papers/>

HULL, D. A., & GREFFENSTETTE, G. 1996. Querying Across Languages: A Dictionary-based Approach to Multilingual Information Retrieval. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval: ACM SIGIR.1996. pp. ??

<http://www.xerox.fr/people/grenoble/hull/papers/sigir96.ps>

HULL, D. A. 1997. Using Structured Queries for Disambiguation in Cross-Language Information Retrieval, AAAI Symposium on Cross-Language Text and Speech Retrieval: American Association for Artificial Intelligence.1997. pp. 84-98. ISBN: 1-57735-040-5; Technical Report: SS-97-05. <http://www.clis.umd.edu/dlrg/filter/sss/papers/>

HUTCHINS, W. J., PARGETER, L. J., & SAUNDERS, W. L. 1971. The Language Barrier. Sheffield: University of Sheffield Postgraduate School of Librarianship and Information Science

ILJON, A. 1977. Scientific and technical data bases in a multilingual society. On-Line Review, 1977; 1(2): 133-136

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO). 1985. Guidelines for the establishment and development of multilingual thesauri: ISO English version. ISO 5964-1985 (E) distributed by the American National Standards Institute.

INTERNATIONAL TRANSLATIONS CENTRE. 1996. Word Translations Index. Delft, The Netherlands, International Translations Centre. Vol. 10, #9. ISSN 0259-8264.

INTERNET SOCIETY & ALIS TECHNOLOGIES. 1997. Web Languages Hit Parade. <http://www.isoc.org:8080/palmares.en.html>

JONES, G. J. F., & JAMES, D. A. 1997. A Critical Review of State-of-the-Art Technologies for Cross-Language Speech Retrieval, AAAI Symposium on Cross-Language Text and Speech Retrieval: American Association for Artificial Intelligence.1997. pp. 99-110. ISBN: 1-57735-040-5; Technical Report: SS-97-05. <http://www.clis.umd.edu/dlrg/filter/sss/papers/>

KALACHKINA, S. Y. 1987. Algorithmic Determination of Descriptor Equivalents in Different Natural Languages. Automatic Documentation and Mathematical Linguistics, 1987; 21(4): 21--29. English translation from Russian.

KARATZOGLOU, M. 1997. TRANSLIB (LIB/3-3038). Patras, Greece: Knowledge S.A.

KIKUI, G. 1996. Identifying the Coding System and Language of On-line Documents on the Internet. In: Sixteenth International Conference on Computational Linguistics (COLING). International Committee on Computational Linguistics. <http://isserv.tas.ntt.jp/chisho/paper/9608KikuiCOLING.ps>

KIKUI, G., HAYASHI, Y. & SUZAKI, S. 1996. Cross-lingual Information Retrieval on the WWW. In: Proceedings of the First Workshop on Multilinguality in Software Engineering: The AI Contribution (MULSAIC). European Coordinating Committee for Artificial Intelligence.

<http://isserv.tas.ntt.jp/chisho/paper/9608KikuiMULSAIC.ps.Z>

KRAAIJ, W. 1997. Multilingual Functionality in the TwentyOne Project, AAAI Symposium on Cross-Language Text and Speech Retrieval: American Association for Artificial Intelligence.1997. pp. 127-132. ISBN: 1-57735-040-5; Technical Report: SS-97-05.

<http://www.clis.umd.edu/dlrg/filter/sss/papers/>

KRAAIJ, W. & HIEMSTRA, D. TREC6 Working Notes: Baseline Tests for Cross Language Retrieval with the Twenty-One System. In: TREC6 working notes. National Institute of Standards and Technology (NIST), Gaithersburg, MD.

KROVETZ, R. & CROFT, B. 1992. Lexical Ambiguity and Information Retrieval. ACM Transactions on Information Systems, Vol. 10, No. 2, pp. 115-141.

KWOK, K. L. 1997. Evaluation of an English-Chinese Cross-Lingual Retrieval Experiment, AAAI Symposium on Cross Language Text and Speech Retrieval : American Association for Artificial Intelligence.1997. pp. 133-137. ISBN: 1-57735-040-5; Technical Report: SS-97-05.

<http://www.clis.umd.edu/dlrg/filter/sss/papers/>

LANDAUER, T. K., & LITTMAN, M. L. 1990. Fully Automatic Cross-Language Document Retrieval Using Latent Semantic Indexing, Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research Waterloo, Ontario: UW Centre for the New OED and Text Research.1990. pp. 31--38.

<http://www.cs.duke.edu/~mlittman/docs/x-lang.ps>

LANDAUER, T. K., & LITTMAN, M. L. 1991. A statistical method for language-independent representation of the topical content of text segments, Proceedings of the Eleventh International Conference: Expert Systems and Their Applications (Vol. 8) Avignon France.1991. pp. 77--85.

LEBOWITZ, A. I., ZWART, R. P., & SCHMID, H. 1991. Multilingual Indexing and Retrieval in Bibliographic Systems: The AGRIS Experience. Quarterly Bulletin of the International Association of Agricultural Librarians and Documentalists, 1991; 36(3): 187-192

LEE, D., NOHL, C. R. & BAIRD, H. ? Language Identification in Complex, Unoriented, and Degraded Document Images.

- LI, C. S., POLLITT, A. S., & SMITH, M. P. 1992. Multilingual MenUSE - A Japanese front-end for searching English Language databases and vice versa. In: T. McEnery & C. Paice (Eds.), 14th Information Retrieval Colloquium. New York: Springer-Verlag. 1992. pp. ?? ISBN: 3540198083, 0387198083.
- LIN, C.-H., & CHEN, H. 1996. An Automatic Indexing and Neural Network Approach to Concept Retrieval and Classification of Multilingual (Chinese-English) Documents. IEEE Transactions on Systems Man and Cybernetics, 1996; 26(1): 75--88.
<http://ai.bpa.arizona.edu/papers/chinese93/chinese93.html>.
- LOGINOV, B. R., & V'YUGIN, V. V. 1989. Automated Maintenance of a Bilingual Medical Thesaurus on a Microcomputer. Automatic Documentation and Mathematical Linguistics, 1989; 23(2): 72--75. English translation from Russian.
- LOUKACHEVITCH, N. V. 1997. Knowledge Representation for Multilingual Text Categorization, AAAI Symposium on Cross-Language Text and Speech Retrieval : American Association for Artificial Intelligence. 1997. pp. 138-142. ISBN: 1-57735-040-5; Technical Report: SS-97-05. <http://www.clis.umd.edu/dlrg/filter/sss/papers/>
- MALONE, T. W., Grant, K. R., Turbak, F. A., Brobst, S. A., & COHEN, M. D. 1987. Intelligent Information Sharing Systems. Communications of the ACM. Vol. 30, no. 5, pp. 390-402.
- MARCHIONINI, G. 1995. Information Seeking in Electronic Environments. Cambridge University Press.
- MATEEV, B., MUNTEANU, E., SHERIDAN, P., WECHSLER, M., and SCHÄUBLE, P. 1998. ETH TREC-6: Routing, Chinese, Cross-Language and Spoken Document Retrieval. In: Proceedings of the Sixth Text Retrieval Conference (TREC-6), National Institute of Standards and Technology (NIST), Gaithersburg, MD.
- MEADOWS, A. J. 1974. Communication in Science. London: Butterworths
- METOYER-DURAN, CHERYL. 1993. Information Gatekeepers. In: Annual Review of Information Science and Technology. Medford, NJ: American Society for Information Science, pp. 111-150. Vol. 28, chapter 3.
- MILLER, G. 1990. WordNet: An On-line Lexical Database. International Journal of Lexicography, Vol. 3, no. 4. Special Issue.
- NELSON, P. 1991. Breaching the Language Barrier: Experimentation with Japanese to English Machine Translation. In: D. I. Raitt (Ed.), 15th International Online Information Meeting Proceedings: Learned Information. 1991. pp. 21--33.
- NEVILLE, H. H. 1970. Feasibility study of a scheme for reconciling thesauri covering a

common subject. *Journal of Documentation*, 1970; 26(4): 313--336.

NEVILLE, H. H. 1975. Alternatives to conventional multilingual thesauri (British Library Research and Development Report 5265 HC)

NG, K. & ZUE, V. W. Phonetic Recognition for Spoken Document Retrieval. Proceedings of IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 1998.

NGUYEN, V. B. H., WILKINSON, R., & ZOBEL, J. 1997. Cross-Language Retrieval in English and Vietnamese, AAI Symposium on Cross-Language Text and Speech Retrieval : American Association of Artificial Intelligence.1997. pp. 143-145. ISBN: 1-57735-040-5; Technical Report: SS-97-05.
<http://www.clis.umd.edu/dlrg/filter/sss/papers/>

OARD, D. W. 1996. Adaptive Vector Space Text Filtering for Monolingual and Cross-Language Applications. Unpublished PhD Dissertation, University of Maryland, College Park.

OARD, D. W. 1997a. Adaptive Filtering of Multilingual Document Streams. In: Fifth RIAO Conference on Computer Assisted Information Searching on the Internet. 1997. pp. ?
<http://www.glue.umd.edu/dlrg/~oard/research.html>

OARD, D. W. 1997b. Alternative Approaches for Cross-Language Text Retrieval, AAI Symposium on Cross-Language Text and Speech Retrieval : American Association for Artificial Intelligence.1997. pp. 154-162. ISBN: 1-57735-040-5; Technical Report: SS-97-05.
<http://www.glue.umd.edu/~oard/research.html>

OARD, D. W. 1997c. Serving Users in Many Languages : Cross-Language Information Retrieval for Digital Libraries. D-Lib Magazine. Vol. ?, No? Dec .
<http://www.dlib.org>

OARD, D. W., & DORR, B. J. 1996. A Survey of Multilingual Text Retrieval (CS-TR-3615): University of Maryland, Institute for Advanced Computer Studies.
<http://www.glue.umd.edu/~oard/research.html>

OARD, D. W., & DORR, B. J. 1998. Evaluating Cross-Language Text Filtering Effectiveness. In: G. Grefenstette (Ed.), Cross Language Information Retrieval: Kluwer Academic. pp. ?. ISBN:0-7923-8122-X.
<http://www.glue.umd.edu/~oard/research.html>

OARD, D. W., DORR, B. J., HACKETT, P. G., & KATSOVA, M. 1998. A Comparative Study of Knowledge-Based Approaches for Cross-Language Information Retrieval. Institute for Advanced Computer Studies, University of Maryland. CS-TR-3897.

OARD, D. W. & HACKETT, P. 1998. Document Translation for Cross-Language Text Retrieval at the University of Maryland. In: Proceedings of the Sixth Text Retrieval Conference (TREC-6), National Institute of Standards and Technology (NIST), Gaithersburg, MD. <http://trec.nist.gov>

OFFICE FOR OFFICIAL PUBLICATIONS OF THE EUROPEAN COMMUNITIES. 1995. Thesaurus EUROVOC Volume 3: Multilingual version. Luxembourg.

O'HAFAN, M. 1996. The Coming Industry of Teletranslation. Clevedon: Multilingual Matters. ISBN 1-85359-326-5.

PASANEN-TUOMAINEN, I. 1991. Analysis of Subject Searching in the TENTTU Books Database. In: J. K. Lucker (Ed.), Proceedings of the 14th Biennial Conference of IATUL (Vol. 1): International Association of Technological University Libraries. 1991. pp. 72--77.

PASHCHENKO, N. A., KALACHKINA, S. Y., MATSAK, N. M., & FIGUR, V. A. 1982. Basic Principles for Creating Multilanguage Information Retrieval Thesauri (Experience with implementing GOST 7.24-80). Automatic Documentation and Mathematical Linguistics, 1982; 16(3): 30--36. English translation from Russian.

PELLISSIER, D., & ARTUR, O. 1986. The Multilingual Evolution of PASCAL, 10th International Online Information Meeting : Learned Information. 1986. pp. 113--121.

PETERS, C., & PICCHI, E. 1997. Using Linguistic Tools and Resources in Cross-Language Retrieval, AAAI Symposium on Cross-Language Text and Speech Retrieval : American Association for Artificial Intelligence. 1997. pp. 179-188. ISBN: 1-57735-040-5; Technical Report: SS-97-05. <http://www.clis.umd.edu/dlrg/filter/sss/papers/>

PEVZNER, B. R. 1969. Automatic Translation of English Text to the Language of the Pusto- Nepusto-2 System. Automatic Documentation and Mathematical Linguistics, 1969; 3(4): 40--48. English translation from Russian.

PEVZNER, B. R. 1972. Comparative Evaluation of the Operation of the Russian and English Variants of the "Pusto- Nepusto-2" System. Automatic Documentation and Mathematical Linguistics, 1972; 6(2): 71--74. English translation from Russian.

PICCHI, E., & PETERS, C. 1996. Cross Language Information Retrieval: A System for Comparable Corpus Querying. In: G. Grefenstette, A. Smeaton, & P. Sheridan (Eds.), Workshop on Cross-Linguistic Information Retrieval : ACM SIGIR. 1996. pp. 24--33. <http://www.rsrc.xerox.com/research/mltt/DMHead/CLIR/>

FIGUR, V. A. 1979. Multilanguage Information-Retrieval Systems: Integration Levels and Language Support. Automatic Documentation and Mathematical Linguistics, 1979; 13(1): 36--46. English translation from Russian.

PIONEER CONSULTING 1997. Pioneer Forecast: International E-mail Growth. The Pioneer Report, 1997; 1(aug): 3.

<http://www.pionerconsutling.com>

POLLITT, A. S., ELLIS, G. P., SMITH, M. P., GREGORY, M. R., LI, C. S., & ZANGENBERG, H. 1993. A Common Query Interface for Multilingual Document Retrieval from Databases of the European Community Institutions. In: D. I. Raitt & B. Jeapes (Eds.), 17th International Meeting on Online Information : Learned Information.1993. pp. 47--61. ISBN: 0904933857

POLLITT, A. S., & ELLIS, G. P. 1993. Multilingual access to document databases, 21st Annual Conference Canadian Society of Information Science .1993. pp. 128-140.

RADWAN, K. 1994. Vers l'Accès Multilingue en Langage Naturel aux Bases de Données Textuelles. Unpublished PhD, Université de Paris-Sud, Centre d'Orsay.

RADWAN, K., & FLUHR, C. 1995. Textual database lexicon used as a filter to resolve semantic ambiguity applications on multilingual information retrieval, 4th Annual Symposium on Document Analysis and Information Retrieval: University of Nevada.1995. pp. 121-136.

READWARE

<http://?>

REHDER, B., LITTMAN, M., DUMAIS, S., & LANDAUER, T. 1998. Automatic 3-Language Cross-Language Information Retrieval with Latent Semantic Indexing. In: Proceedings of the Sixth Text Retrieval Conference (TREC-6), National Institute of Standards and Technology (NIST), Gaithersburg, MD.

RESNIK, P. 1997. Evaluating Multilingual Gisting of Web Pages, AAAI Symposium on Cross-Language Text and Speech Retrieval : American Association for Artificial Intelligence.1997. pp. 189-195. ISBN: 1-57735-040-5; Technical Report: SS-97-05.

<http://www.clis.umd.edu/dlrg/filter/sss/papers/>

RIDDLE, J. N. 1992. FBIS Requirements and Capabilities. In: First International Symposium on National Security and National Competitiveness. Open Source Solutions (OSS), pp. 264-271. <http://www.oss.net>

RIGBY, M. 1981. Automation and the UDC 1948--1980. (2 ed.). The Hague: Federation Internationale de Documentation (FID)

ROLLAND-THOMAS, P., & MERCURE, G. E. 1989. Subject Access in a Bilingual Online Catalog. Cataloging and Classification Quarterly, 1989; 10(1/2): 141--163

ROLLING, L. 1975. Multilingual systems: Survey of the European scene. In: V.

Horsnell (Ed.), Report of a Workshop on Multilingual Systems.1975. pp. 4--5. British Library Research and Development Report 5265 HC

SALTON, G. 1970. Automatic Processing of Foreign Language Documents. Journal of the American Society for Information Science, 1970; 21(3): 187--194

SALTON, G. 1973. Experiments in Multi-Lingual Information Retrieval. Information Processing Letters, 1973; 2(1): 6--11. TR 72-154. <http://cs-tr.cs.cornell.edu>

SANDERSON, M. 1994. Word Sense Disambiguation and Information Retrieval. In: Croft, B. & van Rijsbergen, K. (Eds.), Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. pp. 142-151. Springer-Verlag.
<http://www.dcs.gla.ac.uk/ir/papers/Postscript/sanderson94b.ps.gz>

SCHÄUBLE, P. & SHERIDAN, P. (1998) Cross-Language Information Retrieval (CLIR) Track Overview. In: Proceedings of the Sixth Text REtrieval Conference (TREC-6), National Institute of Standards and Technology (NIST), Gaithersburg, MD.

SEMTURS, F. 1978. STAIRS/TLS - A System for "Free Text" and "Descriptor" Searching. In: E. H. Brenner (Ed.) Vol. 15: American Society for Information Science.1978. pp. 295--298.

SHERIDAN, P. & BALLERINI, J. P. 1996. Experiments in Multilingual Information Retrieval Using the SPIDER System. In: H. P. Frei (Ed.), Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 58-66. ISBN: 0897917928, 3891919999. <http://www-ir.inf.ethz.ch/Public-Web-Pages/sheridan/papers/SIGIR96.ps>

SHERIDAN, P, BALLERINI, J. P., & SCHÄUBLE, P. 1998. Building a Large Multilingual Test Collection from Comparable News Documents. In: G. Grefenstette, (Ed.): Cross Language Information Retrieval. Kluwer Academic. pp. ?? . ISBN:0-7923-8122-X. <http://www.rxrc.xerox.com/research/mltt/DMHead/CLIR/>

SHERIDAN, P., & SCHÄUBLE, P. 1997. Cross-Language Information Retrieval in a Multilingual Legal Domain, Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries, Pisa, Italy. pp. 253-268.
<http://www-ir.inf.ethz.ch/Public-Web-Pages/sheridan/papers/sheridan.html>

SHERIDAN, P., WECHSLER, M., & SCHÄUBLE, P. 1997. Cross-Language Speech Retrieval: Establishing a Baseline Performance. In: N. J. Belkin, A. D. Narasimhalu, & P. Willett (Eds.), Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval . pp. 99-109. ISBN: 0897918363
<http://www-ir.inf.ethz.ch/Public-Web-Pages/sheridan/>

SMEATON, A. F. & SPITZ, A. L. 1997. Using Character Shape Coding for Information Retrieval. In: Proceedings of the Fourth International Conference on Document Analysis and Recognition, ICDAR'97, Ulm, Germany, IEEE Computer Society, pp.974-978.
<http://www.compapp.dcu.ie/~asmeaton/pubs-list.html>

SMITH, M. P. & POLLITT, A. S. 1992. An Evaluation of Concept Translation Through Menu Navigation in the MenUSE Intermediary System. In: McEnery, T. & Pais, C. (Eds.), Proceedings of 14th Information Retrieval Colloquium (BCS). University of Lancaster, pp. 38-54.

SOERGEL, D. 1997. Multilingual Thesauri in Cross-Language Text and Speech Retrieval. In: AAAI Symposium on Cross-Language Text and Speech Retrieval : American Association for Artificial Intelligence.1997. pp. 197-216. ISBN: 1-57735-040-5; Technical Report: SS-97-05. <http://www.clis.umd.edu/dlrg/filter/sss/papers/>

STAMATATOS, E., MICHOS, S., PATELODIMOU, C., & FAKOTAKIS, N. 1997. TRANSLIB : An Advanced Tool for Supporting Multilingual Access to Library Catalogues. In: Second Workshop on Multilinguality in the Software Industry: The AI Contribution: International Joint Conference on Artificial Intelligence.1997. pp. <http://www.iit.nrcps.ariadne-t.gr/~costass/mulsaic97.html>

STEGENTRITT, E. 1994. German Analysis: Morpho-Syntax Within the Framework of the Free-Text Retrieval Project E.M.I.R. Saarbrücken, Germany: AQ-Verlag

STUDEMANN, W. 1992. Teaching the Giant to Dance: Contradictions and Opportunities in Open Source within the Intelligence Community. In: Proceedings of the First International Symposium on National Security and National Competitiveness. pp. 82-92 dec. Vol. 2. <http://www.oss.net>

SUZUKI, M., & HASHIMOTO, K. 1996. Enhancing Source Text for WWW Distribution. In: S. H. Myaeng (Ed.), Proceedings of the Workshop on Information Retrieval with Oriental Languages : Korea Research & Development Information Center. 1996. pp. 51--56.

SYNELIS, C. 1995. TRANSLIB User Survey Report (TRANSLIB Technical Report): University of Patras Central Library

TAYLOR, R. S. 1962 The Process of Asking Questions. American Documentation, Vol. 13, no. 4, pp. 391-396.

UNESCO 1971. Guidelines for Establishment and Development of Multilingual Scientific and Technical Thesauri for Information Retrieval. Paris, France: UNESCO report number: SC/WS/501.

VOLODIN, K. I., GUL'NITSKII, L. L., MAKSAKOVA, R. N., PARKHOMENKO, V.

- F., POZHARISKII, I. F., FEDOTOVA, L. V., & YAKOVLEVA, N. I. 1991. Bilingual Indexing of Geological Documents. *Automatic Documentation and Mathematical Linguistics*, 1991; 25(6): 43--45. English translation from Russian.
- WECHSLER, M., SHERIDAN, P., & SCHÄUBLE, P. 1997. Multi-Language Text Indexing for Internet Retrieval. In: Fifth RIAO Conference on Computer-Assisted Information Searching on the Internet. pp. ??
<http://www-ir.inf.ethz.ch/Public-Web-Pages/sheridan/>
- WEIGAND, H. 1997. A Multilingual Ontology-based Lexicon for News Filtering --- The TREVI Project, IJCAI Workshop on Ontologies and Multilingual NLP : International Joint Conference on Artificial Intelligence.1997. pp. ?? <http://crl.nmsu.edu/Events/IJCAI/>
- WELLISCH, H. 1973. Linguistic and Semantic Problems in the Use of English-Language Information Services in Non-English-Speaking Countries, *International Library Review*, vol. 5, no. 2, pp. 147-162.
- WHITNEY, G. 1990. *Language Distribution in Databases: An Analysis and Evaluation*, Metuchen, NJ: Scarecrow Press. ISBN 0-8108-2323-3.
- WILKENSON, R. 1997. Chinese Document Retrieval at TREC-6. In: Harman, D. K. (Ed.) *The Sixth Text Retrieval Conference (TREC-6)*. National Institute of Standards and Technology (NIST), Gaithersburg, MD.
http://trec.nist.gov/pubs/trec6/t6_proceedings.html
- WOOD, D. N. 1967. The Foreign-Language Problem Facing Scientists and Technologists in the United Kingdom : Report of a Recent Survey. *Journal of Documentation*. Vol. 23, no. 2, p. 117-130.
- WOOD, D. N. 1974. Access to Information in Foreign Languages -- An Experiment. *BLL Review*, Vol. 2, #1, pp. 12-14.
- YAMABANA, K., MURAKI, K., DOI, S., & KAMEI, S.-I. 1998. A Language Conversion Front-End for Cross-Linguistic Information Retrieval. In: G. Grefenstette (Ed.), *Cross Language Information Retrieval*: Kluwer Academic. pp. ??. ISBN:0-7923-8122-X. <http://www.rxrc.xerox.com/research/mltt/DMHead/CLIR/>
- YANG, Y., BROWN, R. D., FREDERKING, R. E., CARBONELL, J. G., GENG, Y., & LEE, D. 1997. Bilingual-corpus Based Approaches to Translingual Information Retrieval, Second Workshop on Multilinguality in the Software Industry: The AI Contribution: International Joint Conference on Artificial Intelligence.1997. pp. ??
<http://www.iit.nrcps.ariadne-t.gr/~costass/mulsaic97.html>

ZISSMAN, M. A. 1996. Comparison of Four Approaches to Automatic Language Identification of Telephone Speech, IEEE Trans. Speech and Audio Proc., SAP-4(1), pp. 31-44.