

[This chapter is available as <http://www.cs.cmu.edu/~ref/mlim/chapter2.html> .]

[Please send any comments to Robert Frederking ([ref+@cs.cmu.edu](mailto:ref+@cs.cmu.edu), Web document maintainer) or [Ed Hovy](#) or [Nancy Ide](#).]

## Chapter 2

### Multilingual (or Cross-lingual) Information Retrieval

**Editors: Judith Klavans and Eduard Hovy**

**Contributors:**

Christian Fluhr

Robert E. Frederking

Doug Oard

Akitoshi Okumura, Kai Ishikawa, and Kenji Satoh

#### Abstract

The term Multilingual Information Retrieval (MLIR) involves the study of systems that accept queries for information in various languages and return objects (text, and other media) of various languages, translated into the user's language. The rapid growth and online availability of information in many languages has made this a highly relevant field of research within the broad umbrella of language processing research. We ignore here issues pertaining to Machine Translation (Chapter 4) and Multimedia (Chapter 9), and focus on the extensions required of traditional Information Retrieval (IR) to handle more than one language.

#### 2.1 Multilingual Information Retrieval

##### 2.1.1 Definition and Terms

Multilingual Information Retrieval (MLIR) refers to the ability to process a query for information in any language, search a collection of objects, including text, images, sound files, etc., and return the most relevant objects, translated if necessary into the user's language. The explosion in recent years of freely-distributed unstructured information in all media, most notably on the World Wide Web, has opened the traditional field of Information Retrieval (IR) up to include image, video, speech, and other media, and has extended out to include access across multiple languages. Being new, MLIR will probably also include the historically excluded access mechanisms typical of libraries involving structured data, such as MARC catalogue records.

The general field of MLIR has expanded in several directions, focusing on different issues; what exactly is within its purview remains open to discussion. It is generally agreed, however, that Machine Translation proper (see [Chapter 4](#)) and Multimedia processing (see [Chapter 9](#)) are not included. Nonetheless, several new terms have arisen around the new IR, each with a slight variation in emphasis, inclusiveness, or historical association with related fields. For example, recent research in multilingual information retrieval, such as (Fluhr et al., 1998) in (Grefenstette, 1998), includes descriptive catalogue data from libraries as well as unstructured data. Hull and Grefenstette (1996) list five uses of the term MLIR:

1. Monolingual IR in any language other than English. This was the usage from the TREC conference series (Harman 1995) in which IR experiments in Spanish and other languages are referred to as the *multilingual track*.
2. IR performed on a collection of documents in various languages, the documents parallel (paired across languages) or not, with queries entered in one language only. In this case, typically the query is translated and each language-specific portion of the multilingual collection is treated as a separate monolingual section.
3. IR on a monolingual document collection that can be queried in multiple languages. The query is entered in more than one language and typically translated into the document language.
4. IR on a multilingual document collection over which queries in various languages can retrieve documents in various languages. This is an extension of (2) and (3).
5. IR on individually multilingual documents, where more than one language may be present in a single document. This rather curious case may occur when an original language quote is embedded within a document in a different language.

In addition to MLIR, four related terms have been used:

1. Multilingual Information Access (MLIA). The broadest possible term to use is Multilingual Information Access, which refers to query, retrieval, and presentation of information in any language. The term MLIA is used in the NSF-EU working groups (Klavans and Schäuble, 1998). In general, the use of information *access* rather than *retrieval* implies a more general set of access functions, including those that have been part of the traditional library, as well as other modalities of access to other media. Access could refer to the use of speech input for video output, where the language component could consist of close-captioned text or text from speech recognition, or catalogue querying to metadata. The term *information access* came into use recently as a way to broaden the historically narrower use of information retrieval.
2. Multilingual Information Retrieval (MLIR). This term refers to the ability to process a query in any language and return objects, such as text, images, sound files, etc., relevant to the user query in any language. Historically, however, Information Retrieval (IR) as a field involved a group of researchers from the unstructured text data base community who employed statistical methods to match query and document (Salton, 1988). In general, this work was English dominated, given the amount of digital information made available to the research community in the early years in English, and excluded access mechanisms typical of libraries involving structured data, such as MARC catalogue records. Thus MLIR as used in this chapter denotes a significantly wider field of interest than that of traditional IR.
3. Cross-lingual Information Access. The use of the term cross-lingual refers (in this context) to bridging two languages, rather than the ability to access information in any language starting with input any language. Systems with cross-lingual capability can accept a query in language L<sub>1</sub> or L<sub>2</sub>, for example English and French, and are capable of returning documents in either L<sub>1</sub> or L<sub>2</sub>. (In other meetings, the term cross-lingual (or translingual) has been used to distinguish systems that cross a language barrier, as opposed to multiple monolingual systems as in TREC.) This term logically includes access via catalogue record and other structured indexing, as for MLIA.
4. Cross-lingual Information Retrieval (CLIR). CLIR generally implies a relationship to IR, with all the implications that apply to MLIR. At the [1997 Cross-language Information Retrieval Spring Symposium of the American Association of Artificial Intelligence \(Oard et al., 1997\)](#), CLIR was defined with the following research challenge: Given a query in any medium and any language, select relevant items from a multilingual multimedia collection which can be in any medium and any language, and present them in the style or order most likely to be useful to the user, with identical or near-identical objects in different media or languages appropriately identified. This definition of the requirements of a system gives full recognition to the query, retrieval, presentation requirements of a working system from a user perspective, and encapsulates succinctly the full set of capabilities to be included. However, its breadth makes it fit well with a definition of MLIA, the most general term, rather than CLIR, a more precise term.

##### 2.1.2 MLIR: Linking and Hybridizing IR and MT

Multilingual Information Retrieval is a hybrid subject area, interacting with or encompassing several other fields. Section 2.5 discusses related fields.

##### How MLIR Relates to Information Retrieval

MLIR is an application of information retrieval. In many respects, as discussed above, the two fields share exactly the same goals; as such, well-known IR techniques such as vector space indexing, latent semantic indexing (LSI), similarity functions for matching documents, and query processing procedures are equally useful in MLIR. However MLIR differs from IR in several significant ways. Most important, IR involves no translation component, since only one language is involved. The related but not identical problems of translating queries and documents are discussed below. Subsidiary problems, such as keeping track of translations across several languages, are also not part of the standard monolingual information retrieval process.

##### How MLIR Relates to and Uses Machine Translation

The goal in machine translation (MT; see [Chapter 4](#)) is to convert a text, written in language L<sub>1</sub>, into a coherent and accurate translation in language L<sub>2</sub>. To do so, most MT systems convert the input text, usually sentence by sentence, into a series of progressively more abstract internal representations, in which sentence-internal relationships are determined and the intended meaning of each word

is identified. Armed with this information, the appropriate conversions are made to support the output language, upon which output realization, usually also sentence by sentence, is performed. MT requires that the meaning of each individual word be known (as does accurate IR); without this knowledge, homographs (for example *plane*, which can refer to an airplane, carpentry tool, geometric surface, the action of skimming over water, and several other meanings) cannot be translated into their intended foreign words. Without word translation, no output is possible.

### Can MLIR be Achieved by Coupling IR and MT?

Unfortunately, while at first blush it may seem that MLIR is simply a matter of coupling IR and MT engines, the special nature of MLIR places constraints on the input to MT that makes a straightforward coupling infeasible. At one extreme, some recent MLIR research has explored extending IR-based indexing techniques to directly bridge language gaps with no explicit translation step at all; see Sections 2.2.2 and 2.3.1 below. Arguments regarding the special nature of MLIR, contained in the NSF-EU MLIA Working Group White Paper (Klavans and Schäuble, 1998), are summarized here.

Differences between the two types of input submitted by MLIR for translation—queries and documents—necessitate two different types of Machine Translation. In the case of queries, the input to MT is a set of disconnected words, or possibly multi-word phrases. There is no call for MT to parse the input, since no syntactic sentence structure can be found. More seriously, the MT system cannot apply traditional methods of wordsense disambiguation, since the input is not a semantically coherent text. It will have to employ other (possibly IR-like) methods to determine the sense of each polysemous word in order to furnish accurate translations. On the other hand, there is no need to produce a linear, coherent output, and in fact multiple (correct) translations of a query term can provide a form of query expansion, which can improve IR performance. Finally, the processes of sentence planning and sentence realization are irrelevant when the input is a string of isolated query words. Without accurate queries, IR accuracy falls dramatically (results of recent studies are given later in this chapter).

For the stage of IR after retrieval (that is, in the case of retrieved documents), in contrast, documents can be translated back into the user's language using the normal methods of MT. However, also for this part of MLIR, partial translation, or keyword extraction and translation, is often adequate for the user's needs. In particular, given the computational expense of MT, it may be inefficient to translate a full document that the user later determines is not exactly what was desired. In addition, fully general purpose MT (especially between a wide variety of languages) is a very difficult problem. Translating a few keywords or a summary (see [Chapter 3](#)) is often a wise policy.

Several additional differences between monolingual IR and MLIR arise if the user is familiar with more than one language too. In particular, the user interface must provide differential display capabilities to reflect differing language proficiency levels of users. When more than one user receives the results, translation into several languages may have to be provided. Furthermore, depending on the user's level of sophistication, translation of different elements at different stages can be provided to users for a range of information access needs, including keyword translation, term translation, title translation, abstract translation, specific paragraph translation, caption translation, full document translation, etc. Finally, monolingual IR users can also take advantage of the results of MLIR. Simply the knowledge that a particular query will access a certain number of documents in other languages could, in itself, be valuable information, even if translations are not required.

Thus for MLIR much of the typical MT machinery is irrelevant, or at best only partially relevant. The differences with traditional MT mean that MLIR cannot simply employ MT engines as front-end query translators and back-end document translators.

Rather, efficient ways of coupling together the internal processes of IR and MT engines are required, allowing them to employ the results of the other's intermediate results. It is inevitable that second-generation MLIR systems will exhibit some more-than-surface integration of MT and IR modules.

### 2.1.3 Key Technical Issues for MLIR

We discuss three different positions on what are the key problems in MLIR. Grefenstette (1998) focuses on term choice and filtering. Oard (1998) presents user-centered challenges. Finally, Klavans (1999) outlines a two-part view that accommodates system-directed and user-directed research issues.

**Grefenstette (1998)** outlines three problems involving the processing of query terms for MLIR:

- *How can a query term in  $L_1$  be expressed in  $L_2$ ?*

This problem requires knowing how terms map between languages. Since little or no contextual text is present in the query to help with term disambiguation, this involves knowing the full range of choices of translations, not just one possible translation, coupled with an understanding how different domains affect translation possibilities.

- *What mechanisms determine which of the possible translations of text from  $L_1$  to  $L_2$  should be retained?*

The second problem deals with determining how to filter, from all possible choices, which ones should be retained in the current application. Unlike MT, a MLIR system can retain a wider set of possibilities that can later be automatically filtered, depending on the kinds of variants that are permitted. Thus the MLIR system has to balance the amount of inaccurate translations (noise) that degrade results against the amount of processing performed to disambiguate the terms and ensure accuracy.

- *In cases where more than one translation are retained, how can different translation alternatives be weighed?*

Given that it is advisable to retain a set of well-chosen possible terms for the best retrieval performance, a problem new to MLIR arises. The possibility of assigning alternate weights to different translations permits more accurate term choice. For example, in a compound term such as "morphological change", the first word is quite narrow in translation possibilities (e.g., in French, only one translation *la morphologie*) while the second is more general ("change" could be *changement* or *monnaie*). In such cases, more weight could be given to the first word's translation than to the second. This problem is compounded by the fact that some multi-word terms do not decompose, but should be treated as a collocation. Thus, mechanisms for weighing alternatives must consider individual word translation weights as well as multi-word term translation weights.

Grefenstette points out that the first two problems are also found in machine translation, and still require research for fully effective solutions. The third problem is one that clearly distinguishes MLIR from both MT and IR.

**Oard (1998)**, in presentations during the Workshops on MLIR, outlined a historical view of CLIR that is user-centered in nature. He views the overall problem of CLIR as a series of processes, including query formulation and document selection, involving feedback from system to user and from user to system. The system-internal processes of indexing, document processing, and matching are treated as components supporting direct user interaction. He presents three points of historical perspective:

- Focus since 1990. The primary areas of focus within the past decade have been query formulation, selection, examination, and delivery.
- Exploiting CLIR: Translation on Demand. In later years, an additional focus was placed on the matching process, which relies, in turn, on more attention being paid to particular document processing and indexing processes required for CLIR.
- An Emerging Focus: User Interaction. Finally, the most recent focus is on better matching and delivery of documents to users. This involves further refinement of processing techniques for multilingual documents.

Oard's five challenges for the next five years are given in Section 2.4 below.

**Klavans (1999)** approaches the central problems in a somewhat different way, focusing on two sets of issues. One set involves three questions relating to the parts of the query-retrieval process, and the other set relates to user needs.

*System issues.* If the query-retrieval process is considered in sequential terms, the first task is to process a query, the second is to index documents and information in a way that permits access by a query, and the third is to match and rank the similarity between query and document set in order to choose relevant documents. (This model of IR applies to the traditional vector-based approaches to IR. As discussed in Sections 2.2.2 and 2.3.1, it is rather different for Latent Semantic Indexing (LSI) and related techniques.)

- Query processing. For both standard IR systems and MLIR systems a query is a statement of the information needs of the user which is put to the IR system. The query can be stated as a Boolean expression, as a set of attribute value pairs (for a structured or fielded query), as a set of terms, or as a free form natural language expression. In all cases, operations on queries are a function of the type of query and of the capabilities of the IR system. Queries can be enhanced in a variety of ways, e.g., by term expansion, thesaural expansion, feedback from related terms from successful prior hits, and so on. In MLIR, query processing involves not only the basic parsing and interpretation, but also requires additional processing steps. First, translation can occur at this step. Different techniques for query translation are discussed in Sections 2.2.2 and 2.3.1. Second, determination of constituent elements can involve different modules, depending on query input language. Finally, MLIR offers a wider variety of possible feedback operations.
- Indexing. Document processing forms the core of IR systems, and various alternative operations over documents have been developed. In MLIR, the possible operations are even greater, given the richer internal structure and partitioning of the information collection. Terms and/or documents in different languages can be cross-indexed in various ways, and may even include translations into a common format.
- Matching and ranking. Standard IR systems embody a wide range of similarity functions to choose matches between query and indexed objects and rank the selected information objects. Given the imperfect correspondences between languages (terms, translations, etc.), matching functions for MLIR have even more variability.

*Usability Issues.* IR systems present two main interface challenges: first, how to permit a user to input a query in a natural and intuitive way, and second, how to enable the user to interpret the returned results. A component of the latter encompasses ways to permit a user to comment and provide feedback on results and to iteratively improve and refine results. MLIR brings an added complexity to the standard IR task. Users can have different abilities for different languages, affecting their ability to form queries and interpret results. For example, a user might be proficient in understanding documents in French, but could not produce a query in French. In this case, the user will need to formulate a query in his native language, but will want documents returned only in French, not translated. At the same time, this user may have spotty knowledge of German. In this case, he might request a set of key terms translated to his native language, and not want to view source documents in German at all. Or he may simply want a numerical count, in order to know that for a given query, there are a certain number in German, a certain number in French, a certain number in Vietnamese, and so on. In addition, knowing the specific sources of relevant information may also be very valuable.

Since research and applications in MLIR are so new, a full understanding of user needs has yet to be developed and tested. However, these needs differ from simple MT needs, given the user query

production and refinement stages.

### 2.1.4 Summary of Technical Challenges

MLIR involves at least the following four technical challenges:

- **Indexing:** Should documents be indexed separately by language, or all together? Should the indexed material be aligned, cross-indexed, or independent across languages?
- **Query treatment:** Should disambiguation proceed monolingually or multilingually? Should query term expansion be performed monolingually, multilingually, or both?
- **Cross-language document ranking:** How must documents retrieved in different languages by different retrieval processes be compared? If they contain the same information, how should they be merged?
- **Feedback processing:** How should the user's selection of relevant documents and/or passages be propagated to other languages?

## 2.2 Where We Were Five Years Ago

### 2.2.1 Capabilities Then

The lure of cross language information retrieval attracted experimentation by the IR community early on. Already in 1971, Salton showed that the use of a transfer dictionary for English and French (a bilingual wordlist with predefined mappings between terms) could be used to translate from a query in one language to another (Salton, 1971). This experiment, although ignoring the realistic and challenging problem of ambiguity, nonetheless served the information retrieval community well in providing a model for a viable approach to cross language IR. However, at the same time, the experiment also illustrated some of the exceedingly difficult problems in the language translation and mapping component of a system, namely one to many mappings, gaps in term translations, and ambiguity. Similarly, in a manual test with a small corpus, Pevzner (1972) showed for English and Russian that a controlled thesaurus can be used effectively for query term translation.

For nearly twenty years, the areas of IR and MT remained separate, leaving MLIR somewhat dormant. Apart from a few forays into refining these early techniques, all significant advances in MLIR have been made in the past five years. This is not surprising, given that increased amounts of information are becoming available in electronic format, and the economy is globalizing.

### 2.2.2 Major Methods, Techniques, and Approaches Five Years Ago

We discuss the problem within the framework outlined above.

System issues include the following.

- **Query processing.** Early approaches used created query term correspondence lists between languages  $L_1$  and  $L_2$  by hand. Such transfer dictionaries, incorporating precise translations, eliminated any problem of ambiguity, since terms were manually disambiguated in advance (Salton 1971). Not surprisingly, these systems performed at the same level as monolingual systems. However, the problem of automatic and dynamic query translation across domains remained. In subsequent work, parallel corpora were used to filter multiple senses with some success (Leacock et al., 1993), and recent advances in this area have been even more successful. Among the most creative approaches is Translingual Latent Semantic Indexing (Landauer et al., 1998). The original LSI technique essentially uses singular value decomposition to transform the original vector space into a lower-dimensional vector space, in which it is claimed the new dimensions capture the semantic structure of the original space. The translingual version essentially produces a pair of corresponding LSI transformations for two languages, using a parallel corpus. When a query is processed, it is also transformed into the LSI space, and compared to the documents in this space. Thus this method actually combines query "translation" with the indexing task.
- **Indexing.** Very early research on MLIR paid more heed to careful query translation using thesauri or controlled vocabulary. Performance was then achieved by using the same approach to indexing as was used for monolingual IR.
- **Matching and ranking.** As for indexing, multilingual similarity and ranking metrics relied on the same techniques as used for standard IR, and thus produced analogous performance. The primary point of transfer for the multilingual aspects of IR remained query processing.

Usability issues include the following. Early experiments were performed at such a small scale, more in the nature of proof-of-concept rather than full-fledged large-scale systems. User feedback and user needs were simply not part of what was tested.

### 2.2.3 Major Bottlenecks and Problems Five Years Ago

The three major bottlenecks of the early part of this decade still persist. They are: limited resources for building domain and language models; limited new technologies for coping with size of collections; and limited understanding of the myriad of user needs.

## 2.3 Where We Are Today

The burgeoning field of MLIR field is clearly in evidence, as can be seen in the bibliography in the first major review article on the topic (Oard and Dorr, 1996). Papers cited include related work on machine translation, including some research translated from Russian. There are 16 citations prior to 1980, 10 from 1980-89, and 52 from 1990 to early 1996. The first major book to be published on the topic (Grefenstette, 1998) reflects the same temporal bias. This work is slanted towards IR rather than toward MT. It contains 11 citations prior to 1980, 25 from 1980-89, and 101 from 1990 to very early 1998.

### 2.3.1 Major Methods, Techniques, and Approaches Now

Following the format above, we divide the methods into system-centered and user-centered concerns, although each provides feedback to the other.

System issues include the following:

- **Query processing.** Early approaches relied on manual query expansion, resulting in high quality but labor-intensive translated queries. Clearly, given the explosion in information, such translation is not practical. Research on combining dictionary-based and corpus filtered translation options for query processing is promising (Ballesteros and Croft, 1998). The history of combining corpus and dictionary data for enhanced and expanded machine-readable dictionary (MRD) resources in the MT community is also valuable (Klavans and Tzoukermann, 1996). Since queries are much shorter than documents, focusing attention on query processing is likely to contribute significant results. Another line of current research (Carbonell et al., 1997), inspired by the translingual LSI concept, has sought ways of exploiting parallel corpora to produce novel MLIR techniques that improve on explicit query translation.
- **Indexing.** The standard IR methods of indexing involve a small amount of language-specific processing. Various tasks of multilingual document preparation and preprocessing techniques have been the focus of much recent research. This includes *tokenizing* (for example, for Japanese, separating the continuous character stream into individual words), *part of speech tagging*, *stemming*, and *demorphing* (for example, converting inflected words into their root forms plus associated information, a task that can be quite complex in highly inflected languages such as Arabic). In addition, new techniques for extracting collocational and phrasal information, both monolingually and multilingually, which often rely on "comparable corpora" (as opposed to totally parallel corpora), are being developed (Sheridan et al., 1998).
- **Matching and ranking.** The matching problem for multilingual data is considerably more complex than that for monolingual data. Similarity metrics primarily rely on keyword matching, with some limited thesaural and phrase-based expansion. This is more successful within single domains than across domains, due to the ambiguity problem. For MLIR, similarity metrics not only must cross the boundaries of domain and genre for the monolingual case, but must also cross significant conceptual mismatches for the multilingual case. A simple example is the matching between *eat* in English, which translates into either *essen* or *fressen*, two different verbs in German, depending on whether the subject is human or not. This one-to-many matching also bedevils query term expansion. Statistical approaches using collocations and optimized shallow linguistic analysis approaches show promise. Although it is tempting to endorse deeper linguistically based parsing as a solution, recent work suggests that deeper approaches can contribute only after a first pass using more optimized techniques (see [Chapter 6](#)). Machine learning techniques can also be of help here, since the similarity problem often involves a wide range of parameters that may impact choice.

Usability issues include the following. The development of effective MLIR technology will have no impact if the user's needs and operation patterns are not considered. Since MLIR is a growing field, and since applications are just emerging, formative studies of usability are essential. Currently, there are a limited number of systems in early operation which are providing important data (e.g., EuroSpider, the translate function of AltaVista, multilingual catalogue access). The incorporation of users in the relevance feedback loop is particularly important, since user needs vary greatly. A full review of user needs is found in (Klavans and Schäuble, 1998).

### 2.3.2 Major Bottlenecks and Problems

Since this is a new field, the bottlenecks listed in Section 2.2.3, evident in earlier years, persist.

## 2.4 Where We Will Be in Five Years

The growing amount of multilingual corpora is providing a valuable and as yet untapped resource for MLIR. Such corpora are essential to building successful dynamic term and phrase translation thesauri, which is, in turn, key to effective indexing and matching. One of the key challenges is in devising efficient yet linguistically informed methods of tapping these resources, methods which combine the best of what is known about fast statistical techniques along with more knowledge based symbolic methods. Even promising new techniques, such as translingual LSI (Landauer et al., 1998) and related techniques (Carbonell et al., 1997), will most probably still rely on parallel corpora. Such corpora are often difficult to find, and very expensive to prepare. This has been the motivation for the work on comparable corpora. However, more and more are being created electronically, especially to conform to legal requirements for the European Union. The issues surrounding corpora are extensively discussed in [Chapter 1](#).

An important class of techniques involves machine learning, as applied to the cross-language term mapping problem. Since term translation, loosely defined, is at the core of query processing, document processing, and matching, it is an important process to do thoroughly and accurately. Even if multiple translations are retained in the MLIR process, obtaining a sensible set of domain linked

terms is an important and central task. One way to obtain these term dictionaries is through parallel corpora, but statistical processing is typically difficult to fine tune. As discussed in [Chapter 6](#), machine learning techniques are a fundamental enhancement of the power of language processing systems and hold particular promise in this area as well.

Finally, it is to be hoped that our understanding of user needs and user interactions with MLIR systems will be significantly better in five years than it is now. As early systems emerge and are tested in the field, a range of flexible and fluid applications that can learn and dynamically adjust to the users' levels of competence, across languages and across domains, should appear. One possible example of this type of flexible application might be human-aided MT systems for producing gisting-quality translations of retrieved documents, which would allow the user to make a personal time/quality tradeoff: the longer the user interacted with the translator, the better the resulting output. Most probably, these systems will incorporate multimedia seamlessly and permit multimodal input and output. Such capabilities will provide maximum usability.

#### 2.4.1 Expected Capabilities

Oard (1998) outlines five challenges for the next five years:

- User-assisted query disambiguation, which might be limited to the most troublesome terms;
- Enrichment of dictionary data with unlinked corpora;
- Tailored title translation techniques;
- Rapid translation and/or summarization, which involves some research on using queries to focus the translation effort; and
- Automated global translation brokering, which balances capacity, capability and user needs.

#### 2.4.2 Expected Bottlenecks in Five Years

Four key issues must be overcome in order to achieve effective MLIR. Some of these issues also apply to IR and MT independently.

1. The tension between systems and users. The balance between understanding user needs and building MLIR systems is delicate. On the one hand, applications need to be built in order to test them with users. On the other, users have to define their desiderata for system builders. However, it is difficult to imagine in advance the full set of capabilities that should be part of a MLIR system. Asking system builders or users in advance requires a level of imagination and inventiveness that is difficult to achieve. Therefore a close coupling between these independent but related activities is especially important for building complex MLIR systems.
2. The dependence on resource-expensive technologies. The increased need for multilingual corpora in order to build term translation lists and loose translations in a flexible and domain-independent way brings along an attendant problem: Where will these corpora come from? How reliable are they? Ways to collect, validate, and standardize comparable corpora are needed. Ways to infer associations using other resources and metadata promise some solutions for this problem. Imaginative techniques (for example, using datelines in news articles with proper nouns as anchors, or combining bilingual dictionary data with corpora across languages) will have to be invented.
3. The need for efficiency and accuracy. Different applications require different levels of functionality. In some cases, speed is important and must be prioritized. In others, high precision is a top demand. In others, a wide-ranging glance at the data is all that is needed, so high recall is a more important goal. For each of these priorities, different techniques can be applied. For example, very high precision applications are likely to require more in-depth language analysis, but this type of processing tends to be slow and knowledge intensive. It is important to understand the tradeoffs between shallow statistically motivated techniques and deeper linguistically motivated ones, as discussed in [Chapter 6](#), to achieve processes that are both fast and accurate.
4. The effective presentation of complex information. How should multilingual results of a search be presented back to the user? What kinds of new summarization and visualization techniques will most help people be able to evaluate, digest, and then use the information that is delivered to them? Because multilingual information retrieval adds complexity to the presentation problem, we have yet to fully understand new presentation challenges.

### 2.5 Juxtaposition of This Area with Other Areas

Two major classes of technical issues must be addressed when dealing with multilingual data:

First, technical issues involving data exchange, with a set of attendant sub-issues. This includes questions such as character encoding, font displays, browser/display issues, etc. Such issues have implications for metadata for the Internet, international sharing of bibliographic records, and transliteration and transcription systems.

Second, natural language questions, also with a set of attendant research issues. This includes natural language processing technologies (e.g., syntactic or semantic analysis), machine translation, information retrieval (or information discovery) in multiple languages, speech processing, and summarization. Also included are questions of multilingual language resources, such as dictionaries and thesauri, corpora, and test collections.

The new application of MLIR draws on achievements and techniques in several related areas. However, the challenges unique to MLIR must be handled independently. Listing some of the relevant technologies, these include:

- Information Access: document indexing (multilingual); retrieving, filtering, clustering; presentation and summarization of information; multilingual metadata; cross-language information retrieval. See [Chapter 3](#) and [Chapter 9](#).
- Machine Translation: comparable and parallel text alignment; language generation. See [Chapter 4](#).
- Computational Linguistics: morphological analysis, syntactic parsing, techniques for disambiguation, document segmentation, corpus analysis, creation of derivative lexicons, term recognition and term expansion. See [Chapter 6](#).
- Resources: dictionaries, thesauri, index terms, test collections, speech data bases. See [Chapter 1](#).

Several potentially valuable connections have not yet been made. The Database and Computational Linguistics research and development communities, for example, contain in their members a great deal of relevant expertise. The National Science Foundation PI meeting on Information and Data Management (1998) concluded that closer links between the IR and Database communities would be beneficial to each. Similarly, the human-computer interaction / multimedia community offers important insights into ensuring user-driven design of systems.

In order to facilitate cross-fertilization, a series of small workshops to define new projects, and a series of very small seed projects, would help the specification of prototype systems and the elucidation of complex problem areas. Projects should be interdisciplinary, very limited in scope, with well-defined goals leaving room for exploratory research. The results of such cross-fertilization would depend on the backgrounds of the potential participants. Assembling a group from commerce to assist computer scientists in specifying the needs that MLIR systems must address, or focus groups from high information-needs communities, such as journalism and finance, could be used to specify new projects and prototypes and guide the direction of research in beneficial directions.

### 2.6 References

- Ballesteros, L. and W.B. Croft. 1998. Statistical Methods for Cross-Language Information Retrieval. In G. Grefenstette (ed), *Cross-Language Information Retrieval* (23-40). Boston: Kluwer.
- Carbonell, J., Y. Yang, R. Frederking, R. Brown, Y. Geng, and D. Lee. 1997. Translingual Information Retrieval: A Comparative Evaluation. *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI-97)*. Nagoya, Japan. Best paper award.
- Fluhr, Ch., D. Schmit, Ph. Ortet, F. Elkateb, K. Gurtner, and Kh. Radwan. 1998. Distributed Cross-Language Information Retrieval. In G. Grefenstette (ed), *Cross-Language Information Retrieval* (41-50). Boston: Kluwer.
- Grefenstette, G. (editor) 1998. *Cross-Language Information Retrieval*. Boston: Kluwer.
- Harman, D. (editor) 1995. *Proceedings of the 5<sup>th</sup> Text Retrieval Conference (TREC)*.
- Hull, D. and G. Grefenstette. 1996. Querying across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval. *Proceedings of the 19<sup>th</sup> Annual ACM Conference on Information Retrieval (SIGIR)* (49-57).
- Klavans and Tzoukermann, 1996. Dictionaries and Corpora: Combining Corpus and Machine-readable Dictionary Data for Building Bilingual Lexicons. *Machine Translation* 10 (3-4).
- Klavans, J. and P. Schäuble. 1998. Report on Multilingual Information Access. Report commissioned jointly by NSF and EU.
- Klavans, J. 1999. Work in progress.

- Landauer, T.K, P.W. Foltz, and D. Laham. 1998. An Introduction to Latent Semantic Analysis. *Discourse Processes* 25(2&3) (259-284).
- Leacock, C., G. Towell, and E. Voorhees. 1993. Corpus-Based Statistical Sense Resolution. *Proceedings of the DARPA Human Language Technology Workshop* (260-265). Princeton, NJ.
- Oard, D. and B. Dorr. 1996. A Survey of Multilingual Text Retrieval. Technical Report UMIACS-TR-96-19, University of Maryland Institute for Advanced Computer Studies. <http://www.clis.umd.edu/dlrg/filter/papers/mlir.ps>.
- Oard, D. and B. Dorr. 1998. Evaluating Cross-Language Text Filtering Effectiveness. In G. Grefenstette (ed), *Cross-Language Information Retrieval* (151-162). Boston: Kluwer.
- Oard, D., et al., 1997. Proceedings of the [AAAI Spring Symposium on Cross-Language Information Retrieval](#). San Francisco: Morgan Kaufmann AAAI Press.
- 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* 6(2) (71-74). English translation from Russian.
- Salton, G. 1971. *Automatic Processing of Foreign Language Documents*. Englewood Cliffs, NJ: Prentice-Hall.
- Salton, G. 1988. *Automatic Text Processing*. Reading, MA: Addison-Wesley.
- Sheridan, P., J.P. Ballerini, and P. Schäuble. 1998. Building a Large Multilingual Test Collection from Comparable News Documents. In G. Grefenstette (ed), *Cross-Language Information Retrieval* (137-150). Boston: Kluwer.