

Experiments in Multilingual Information Retrieval

David A. Hull Gregory Grefenstette

Rank Xerox Research Centre
6 chemin de Maupertuis, 38240 Meylan France
{hull,grefen}@grenoble.rxrc.xerox.com

January 31, 1996

Abstract

The multilingual information retrieval system of the future will need to be able to retrieve documents across language boundaries. This extension of the classical IR problem is particularly challenging, as significant resources are required to perform query translation. At Xerox, we are working to build a multilingual IR system and conducting a series of experiments to understand what factors are most important in making the system work. Using translated queries and a bilingual transfer dictionary, we have learned that cross-language multilingual IR is feasible, although performance lags considerably behind the monolingual standard. The experiments suggest that correct identification and translation of multi-word terminology is the single most important source of error in the system, although ambiguity in translation also contributes to poor performance.

1 Introduction

As Internet resources such as the World Wide Web become accessible to more and more countries, and technological advances overcome the network, interface, and computer system differences which have impeded information access, it will become more common for searchers to wish to explore collections of documents that are not written in their native language. Beyond merely accepting extended character sets and performing language identification, the information retrieval systems of the future will have to provide help in searching for information across language boundaries. At Xerox, we have begun a series of experiments to explore what factors are most important in making multilingual information retrieval systems work.

After presenting our definition of multilingual information retrieval, we introduce several basic approaches to the problem and discuss the previous research work in this area. We then describe our first round of experiments using French queries and an English document collection (TIPSTER). We demonstrate that multilingual information retrieval is feasible using a general language bilingual dictionary and some basic linguistic analysis tools, but that there is a significant gap between monolingual and multilingual performance. From a failure analysis of the results, we learn that translation ambiguity and missing terminology are the two primary sources of error, and we conclude by suggesting some methods for resolving these problems.

Our goal in these experiments is not to build the ideal fully-functional multilingual IR system, as the time and resources required for this task are considerable. Rather, we restrict our attention to a single language pair and try to understand the basic requirements for effective multilingual IR and the problems that arise from a simple implementation of such a system. From this research, we can begin to understand which components of the system are most important and find some directions for our future research. Therefore, the examples of when and the reasons why the system failed are more important than the numerical results of the experiments.

2 Defining Multilingual Information Retrieval

There is no common currently accepted definition for multilingual information retrieval (MLIR). The term has been used in the past to cover a broad range of different approaches to information retrieval (IR) using one or more languages. In this section, we will present a number of different descriptions of the multilingual IR task that have been used by previous authors and outline our own approach to the problem. Five different definitions for MLIR are outlined below.

- (1) IR in any language other than English.
- (2) IR on a parallel document collection or on a multilingual document collection where the search space is restricted to the query language.
- (3) IR on a monolingual document collection which can be queried in multiple languages.
- (4) IR on a multilingual document collection, where queries can retrieve documents in multiple languages.
- (5) IR on multilingual documents, i.e. more than one language can be present in the individual documents

Definition (1) comes from the recent TREC conferences [14], where the IR experiments in Spanish are referred to as the multilingual track. A number of modern IR systems claim that they are performing multilingual information retrieval because they are capable of handling extended character sets and performing monolingual IR in multiple languages. An expanded version of this definition is (2), where it is assumed that the system is working with a multilingual document collection but the documents in each language are viewed as separate and independent parts of the collection. Once the query is entered, the language of interest is fixed, and only documents in that language are searched. This definition also covers parallel document collections¹, which can be searched in any of their component languages.

Definition (3) assumes that the document collection is monolingual, but the retrieval system is capable of processing queries in a number of different languages and retrieving documents across language boundaries. With (4), a simple extension of the previous definition, we can move to a multilingual document collection which can be searched in any of its component

¹A parallel document collection is defined as one where the same documents are presented in two or more languages.

languages and where documents can be retrieved in multiple languages in response to a single search. Finally, we can generalize this definition still further to multilingual documents (5). The MLIR problem descriptions above are listed in order of their complexity. Recent work on Spanish IR [2, 5] suggests that modern IR systems and techniques can be applied effectively to other languages. However, the authors feel that real MLIR involves cross-language document retrieval (3-5), and IR research is only beginning to address this important and difficult class of problems.

The MLIR problem as we would like to define it (3-5) requires some form of query translation. This adds an entirely new order of complexity to the traditional IR problem and requires that an extensive knowledge base be incorporated into the system for each query/document language pair which is supported. However, MLIR is a far more tractable problem than machine translation, as the translated representations of documents or queries are for automatic machine use only and need never be read by a human being. Therefore the entire issue of syntactic generation can be avoided. Information retrieval is such an inexact science that it is not clear whether direct query translation is necessary or even optimal for identifying relevant documents. It may be possible to achieve a reasonable level of performance by relying primarily on statistical similarity information. However, recent research by Davis and Dunning suggests that such information alone is not sufficient [10]. This is one of the important issues which needs to be explored in this new field of research.

Our research concentrates on problem (3) above for several reasons. First, in order to obtain experimental results that are both reliable and quantifiable, it is necessary to have test document collection with large numbers of relevance judgements. For the moment, such resources are scarce to non-existent in the multilingual domain. Therefore, we are able to leverage off the extensive previous work in traditional IR if we rely upon a monolingual test collection. Second, following a long tradition of experimental research, we attack the simplest problem first, and worry later about generalizing our results to multilingual document collections.

Definition (4) adds an additional layer of complexity to the IR problem because of the language-dependent nature of the translation process. There is no guarantee that document similarity scores obtained by translation of the query into several different languages will be comparable, so creating a single ranked list of documents requires research into merging strategies that is only beginning to happen for monolingual merging problems [26]. Definition (5) requires that scores be combined on the level of the individual documents. We also restrict ourselves to a single language pair (English/French) in order to minimize the resource requirements of our experiments.

3 Basic Approaches to MLIR

Information Retrieval systems rank documents according to statistical similarity measures based on the cooccurrence of terms in queries and documents. For our approach to MLIR, which tries to capture relevant documents across language boundaries, some mechanism for query or document translation is required. Our MLIR system has been designed to use query translation. There is no reason in principle why the problem could not be approached using document translation instead. Queries tend to be very short (often only one or two words), which does not provide much context for the translation process, so it is quite possible that document translation could lead to better performance. However, when faced with the choice between translating the query once or translating every document in the collection, the former seems much more realistic in terms of efficiency. The query translation module can easily be built on top of an already existing IR software package without having to redesign the entire system. Furthermore, for document translation, the storage costs increase linearly with the number of supported languages (unless document translation is performed dynamically, which is an equally unappetizing option).

Oard et al. [19] define three major approaches to the problem of interlingual term correspondence, which they call text translation, term vector translation, and Latent Semantic Coindexing². Text translation describes the processing of mapping the query from the source language directly into one or more target languages using a machine translation (MT) system. This represents the high-end approach to MLIR, since machine translation generally involves the sophisticated use of natural language processing and text generation techniques. It is characterized by the direct resolution of ambiguity in translation using structural information from the source language text. This strategy for MLIR might allow the researcher to take advantage of the extensive body of research on machine translation and the availability of commercial products. However, the performance of current MT systems in the setting of general language translation is dismal enough to make this option less than entirely satisfactory. Experiments by Radwan [21] confirm these suspicions.

As a more robust alternative, we can consider using term vector translation, a process by which each word in the source language is mapped to all of its possible definitions in the target language. Retrieval strategies based on the vector space model can seamlessly evaluate this extended representation of the translated query. However, this approach raises a number of important issues with respect to term weighting strategies. Should each term be weighted according to the number of translations? For example, a term with four translations may have its importance artificially inflated with respect to a term with a single translation unless this one to many mapping is accounted for in the term weights. Perhaps an extended or weighted Boolean model where all translations of each term are linked by disjunction would be more

²The latter describing Latent Semantic Indexing [18] applied to a parallel document collection.

appropriate. Furthermore, some translations of a term will be much more common than others. Should more common translations be weighted proportionally higher? What resources do we use to obtain this information? These questions suggest that corpus-driven methods for MLIR should be considered.

While direct translation of the query seems like the most viable approach to MLIR, one can also consider methods that derive query translations indirectly using a training corpus. Landauer and Littman [18] and Evans et al. [12] present very compelling techniques based on Latent Semantic Indexing, which uses the singular value decomposition of a parallel document collection to obtain term vector representations which are comparable across all the languages of the collection. Unfortunately, this technique has not yet been seriously tested for the MLIR problem. Davis and Dunning [10] examine several interesting alternative approaches which also use a parallel document collection. For example, one can match the query against sentences in the same language and find terms in matching sentences in other languages that are strongly associated with the query topic. Indirect query translation using training corpora may be able to capture domain and context dependent relationships that would be missed by other techniques. We consider exploiting parallel corpora for direct or indirect translation to be a promising line of research.

4 Resources requirements for MLIR

Extending information retrieval to the multilingual domain requires that researchers obtain a significant number of additional resources, even for the simplest approaches, where the query and the documents returned are both in the same language. The MLIR system must be able to handle the character sets of each language that is supported and multilingual document collections may benefit from some facilities for automatic language recognition [24, 13]. Support for accented character sets may seem simple on the surface, but there are some important issues that must be addressed. In many text collections, accenting is inconsistent, and capitalized letters sometimes lose their accents. Our approach is to handle these problems during the stemming process.

Efficient stemming algorithms for English have been developed over the past twenty-five years, but most of this work will not generalize to other languages. It is not known what investment would be required to create high-performance stemmers for languages other than English. Stemmers are being developed for other languages but most have not been extensively tested. Fortunately, Xerox linguists have developed an alternative solution based on the use of inflectional analysis. This approach is particularly valuable for term translation, as it only conflates word forms which have the same inflectional root, so no additional ambiguity is introduced before translation. In terms of IR performance, it works as well as traditional stemming algorithms in English [17] and provides similar improvements for Spanish [15].

Xerox has developed a methodology [7] which can be used to construct a morphological analyzer for a new language in 8-10 person/months, and tools have already been built for most Western European languages. These analyzers recognize the possible parts of speech of each token in a text and provide a normalized form for a variety of surface forms. For example, the French word *joignaient* (joined) is morphologically analyzed as the 3rd person plural preterite form of *joindre*. The German word *Weingärtnergenossenschaften* is analyzed as the feminine plural noun *Wein#Gärtner#Genosse(n)#schaft* composed of the agglutination of the words *Wein*, *Gärtner*, *Genosse*, and *schaft*. In addition to providing principled stemming, morphological analysis is a necessary step in any subsequent natural language processing, such as noun phrase recognition. It is also crucial for finding term entries in bilingual dictionaries.

Query translation requires extensive resources for each language pair under consideration. Depending on the approach, this may include (1) a machine translation system, (2) bilingual transfer dictionaries, (3) parallel texts, and/or (4) monolingual domain-specific corpora in several languages. Machine translation systems are useful for direct query translation. Bilingual dictionaries are one source of definitions for term vector mapping. Parallel corpora can be used to extract relationships between terms for term vector translation or as a reference for the indirect query translation strategies described in the previous section. Domain-specific monolingual corpora can be an important source for terminology and can be used as a surrogate when parallel corpora are not available in the field of interest, although the process of extracting translations is much more difficult and manual effort may be required.

Bilingual general language dictionaries in machine readable form are more and more available, although they tend to be expensive to obtain. Unfortunately, they are designed with human readers in mind and thus need to be adapted for use by automatic retrieval systems. For term vector translation, the system needs only the direct translations of each entry, which we will define as a bilingual transfer dictionary (also sometimes called a bilingual thesaurus). Bilingual general language dictionaries contain more verbose definitions and examples including large amounts of vocabulary that would not be suitable for IR. Converting a bilingual dictionary to a transfer dictionary is a non-trivial effort.

Parallel text collections can also be used for term vector translation, but doing so accurately requires immense quantities of training text and statistical models of great sophistication. The work of Brown et al. [4] from IBM epitomizes this strategy. They generate not only term translation vectors but corresponding translation probabilities for each link which accurately model the distribution of the training corpus. This approach is extremely compelling but the translation probabilities must be applied with caution, as they may not generalize to other domains and they are generated independent of context. The IBM work is designed for machine translation and has not been tested on the MLIR problem. Much simpler approaches, such as the matching of associated terms from aligned sentences suggested by Davis and Dunning have

not yet been proven to be effective.

Of these two resources, bilingual general language dictionaries are more prevalent than parallel texts of sufficient size to have similar coverage. While creating a transfer dictionary is not a simple process, the effort required to implement the IBM approach makes the former task seem simple in comparison. Therefore, our experiments use a bilingual transfer dictionary. In a sense, one can consider these two approaches as complementary. The transfer dictionary provides broad but shallow coverage of the language. All major words in the language are defined but most technical terminology is missing and translation probabilities are not available. On the other hand, parallel corpora provide narrow but deep coverage, particularly when they concentrate on a single domain. The ideal MLIR system of the future will want to take advantage of both of these resources.

5 Previous Experiments in MLIR

Multilingual IR systems can be evaluated automatically in much the same fashion as monolingual IR systems, using queries with known relevance judgements. However, in the multilingual context, there is a strong desire to compare the results against the optimal performance of the system if the query were perfectly translated. This gives rise to the following approach to evaluation in MLIR. Start with queries, documents, and relevance judgements in a single language. Have the queries translated into another language by human translators. These translated queries are retranslated by the MLIR system, and the results can then be compared to the original queries to get a good sense of the relative performance of the MLIR system. All of the work described below uses some variant of this strategy. We concentrate on experiments concerning the cross-language MLIR problem, as this research is most relevant to our current work.

The first research on MLIR was conducted by Gerald Salton in the late 60's and early 70's [22]. In this work, Salton used two small monolingual collections (468 German and 1095 English abstracts on library science and documentation) and had 48 English queries and an English thesaurus manually translated into German. For example, one thesaurus entry grouped any English word stemming to one of *charge*, *enter*, *entry*, *insert*, or *post* and the German words *eingang*, *eingegangen*, *ingegeben*, *einsatz*, *einstellen*, and *eintragung*. This thesaurus acted as an Interlingua, since any English or German document or query could be represented as a language-independent concatenation of thesaurus entries. Experiments applied English and German versions of the queries to the documents in both languages and found minimal loss in performance due to the translation process.

While the results are impressive, the experiments are based on a small single-domain text collection and extensive manual work is used to construct the multilingual thesaurus. This work

was conducted in the days when the only computationally feasible strategy for text retrieval was to use a controlled vocabulary of index terms. The monolingual baseline from the Salton experiment would probably be far surpassed by modern full-text retrieval systems. Today's researchers cannot afford to manually construct this kind of multilingual thesaurus for text collections with the size and coverage of those currently being used for information retrieval experiments. However, Salton's experiment clearly demonstrates that multilingual IR is feasible with sufficient resources.

Radwan et al. [21] conducted their MLIR experiments on the French-English language pair, translating the 225 queries of the Cranfield collection (1400 English documents on aeronautics) into French. They created both a domain-dependent terminology dictionary and a general language transfer dictionary to aid in query translation and obtained the following results [p. 244]: monolingual system - 0.345, MLIR using term vector translation and transfer dictionaries - 0.270, MLIR using machine translation (SYSTRAN) - 0.215, as measured by average precision at 10-90% recall.

These results provide strong evidence that term vector translation works better than machine translation. As pointed out by the authors, an error in translation can induce irrecoverable silence. It should be noted that the transfer dictionary required extensive manual editing to attain this level of performance and that the multi-word terminology list was also manually constructed. However, SYSTRAN also used special topical dictionaries on aeronautics, space, and military weaponry, so it appears to be a reasonable comparison. We note that the move from controlled vocabulary (Salton's experiment) to full text retrieval has opened a substantial gap between monolingual and cross-language performance. Also, obtaining this level of performance required a significant investment of manpower in the construction of the transfer dictionaries.

Davis and Dunning [10] tried a wide variety of different corpus learning methods on a monolingual Spanish collection (25 queries and 58,000 documents from the Mexican newspaper El Norte) in the context of the TREC-4 experiments. The Spanish queries were translated into English before being processed by the system. Prior to the experiments, the U.N. Corpus (1.6 GB of English-Spanish [also French] parallel text covering activities at the United Nations) was aligned to produce 680,000 sentence pairs to use as training data for their algorithms, which are briefly outlined below:

- (1) Term vector translation using English-Spanish bilingual dictionary
- (2) High frequency Spanish terms from top-ranked English sentences
- (3) Statistically related Spanish terms from top-ranked English sentences
- (4) Optimization of the queries in (2) using Evolutionary Programming
- (5) Translation matrix derived from Singular Value Decomposition

All of the proposed query-translation strategies performed dramatically worse than the monolingual baseline, with term-vector translation (1) working slightly better than the rest. However,

given the time and resource constraints of the TREC environment, it is best to regard these results as preliminary.

While it is perhaps too early to draw conclusions from this research, it appears that corpus learning strategies alone will probably not be sufficient to provide acceptable performance for MLIR. While the TREC Spanish queries are both short and vague, and the U.N. corpus probably covers very different topics than El Norte, this is likely to be the rule more often than it is the exception in multilingual settings. The telling difference between this work and the two previous studies described above is that (as far as we can tell) Davis and Dunning generated their results entirely automatically. No research group currently working on MLIR has yet been able to build an automatic system that performs well without extensive human intervention in the dictionary construction process.

6 The Xerox Experimental Approach

Following the well-motivated and well-tested methodology described in the previous section, we worked with translated French queries and English documents. Our experiments used the TIPSTER text collection and queries 51-100 from the recent TREC experiments [14]. We chose to use only the news component of the collection, which consisted of articles from the Wall Street Journal, AP newswire, and the San Jose Mercury News, and amounted to roughly half a million documents or about 1.6 GB of text. The 50 selected queries were translated into French externally by a professional translator. We used the term vector translation model, with query translations generated by a bilingual transfer dictionary.

A careful examination of the queries and some preliminary experiments made it clear that the original queries were not suitable for multilingual information retrieval due to their length and content. In particular, the *Concept* fields of the queries contain large amounts of specific terminology, much of which has no good translation and was therefore left in English by the human translator. For example, most acronyms (GATT, LBO, OSHA, SALT II, OPEC, FDIC, NRA, LAN) and proper names (Reagan, Bush, Iran-Contra, Toshiba, M-1 Abrams, AH-64 Apache) are not translated and most technical terminology has one unique translation. Using this information, the MLIR system was able to attain unrealistically high levels of performance. While this may be reasonable for a technical domain, we wish to obtain a more general picture of the problems associated with query translation. Researchers have recognized that most real queries are only a few words long, and this is even more likely to be the case when users are not working in their native language. To address this issue, we decided to work with short versions of the queries (average length of seven words) that had been created for previous TREC experiments and translated them into French. While some acronyms and terminology remain, the overwhelming majority of the language independent evidence has been removed.

We also converted an on-line bilingual French \Rightarrow English dictionary (Oxford Hachette, 1994) to a word-based transfer dictionary suitable for text retrieval, which involved the removal of large amounts of excess information. Unfortunately, this was neither a simple nor a fully automatic process. Although the dictionary was marked up in SGML, which allowed for automated filtering from the definition of sections such as pronunciation, etymology and examples, these markings were not always coherent or correct, creating one source of errors in our translation assignments. In some cases it was possible to detect these marking errors. For example, we would filter out for manual treatment any translation containing personal pronouns such as *I* or *we*. 462 of the 34000 definitions were manually treated based on this filter. Sometimes, especially for common words, the dictionary entry was so long and complex that the automatic filtering would get lost, either through bugs in the filter or misplaced SGML markers. For example, automatic extraction of translations of the common French word *prendre* (to take) gave 23 words including *success, break, catch, set, sink, stiffen, take, thicken, find, oneself, lay, idea, bring, charge, handle, pick, client, put, accent, one's, arm, waist, customer*. We left such noisy definitions as they were produced, but a serious manual cleaning needs to be applied. 521 of the transfer dictionary entries (mostly common words) had ten or more translations. Duplicate terms are removed from each definition.

A further and more pernicious error for our system was the use of encyclopedic definitions which elaborate on the translation of the word, introducing contextual words that are not proper translations of the dictionary head word, but rather clues to a human user of how and where the word is used. These clues are often embedded in the heart of the definition and thus reappear associated with the head word after any automatic filtering of the definitions. For example, the definition for the French word 'radiation,' after filtering out predefined fields yielded: *radiation, expulsion, striking off from the register, loss of the license to practice medicine, disbarring*. Filtering out stopwords still left *disbarring, expulsion, radiation, striking, register, loss, license, practice, medicine* as translation equivalents. Ideally, one would only retain *radiation, expulsion,* and perhaps *disbarring*. In an IR setting, it is evident that the terms *register, license,* and *medicine* add considerable noise to the results. Our automatic filtering created an object that more resembles a thesaurus than a dictionary.

The MLIR process consists of three basic steps. First, the query is morphologically analyzed and each term is replaced by its inflectional root. Second, the system looks up each root in the bilingual transfer dictionary and builds a translated query by taking the concatenation of all term translations. Terms which are missing from the transfer dictionary are passed unchanged to the final query. The translated query is then sent to a traditional monolingual IR system. Documents are also normalized to their inflectional roots. This is the simplest possible approach to MLIR, since all issues relating to specialized term weighting and resolving ambiguity in translation are ignored. It is designed to serve as a baseline, and future efforts will improve upon this model.

English: politically motivated civil disturbances

French: troubles civils à caractère politique

Term vector retranslation:

trouble – turmoil discord trouble unrest disturbance disorder

civil – civil civilian courteous

caractère – character nature

politique – political diplomatic politician policy

Table 1: An example of the query translation process for short query Q67

Table 1 shows a sample query.

These experiments use a modified version the SMART information retrieval system [6], which obtains a ranked list of relevant documents by performing similarity calculation according to the vector space model. We apply the following weight functions to each query (q_i) and document (d_j):

$$\text{wt}_{q_{ik}} = \sqrt{qt_{fik}} * \log(N/n_k), \text{ wtd}_{jk} = \sqrt{dt_{fjk}/|d_j|}, \text{ RSV}(q_i, d_j) = \sum_k \text{wt}_{q_{ik}} * \text{wtd}_{jk}$$

where qt_{fik} (dt_{fjk}) is the frequency of term k in q_i (d_j), $\log N/n_k$ is the traditional IDF term weight, and $|d_j|$ is the length of d_j . Documents are ranked in order of their retrieval status value (RSV) and performance is evaluated by comparing the ranked list to known relevance judgments.

The unique characteristics of MLIR suggest specific strategies for evaluation. One would expect in general that translated queries (particularly short ones) will tend to perform worse than the original queries due to errors and ambiguity introduced by the translation process. Success in information retrieval depends on the ability of the user (with help from the system) to find vocabulary which appears in the documents of interest. This task becomes much more difficult when the terminology must cross language boundaries. This suggests that relevance feedback techniques, which improve the query by incorporating information from previously discovered relevant documents will be a particularly important tool in the multilingual setting.

In addition, the user cannot be expected to quickly read and evaluate (or have translated) lots of documents in a foreign language. Therefore, high precision should be an important goal for an MLIR system. Once a few relevant documents have been collected, the system can resort to monolingual relevance feedback to find more relevant documents if high recall is the final goal. Note that for these experiments, we are assuming that the user is working with a monolingual collection. For a multilingual document collection, good relevance feedback would probably necessitate obtaining at least one relevant document in each language of interest. For these reasons, we choose to use precision averaged at 5, 10, 15, and 20 documents retrieved as the evaluation measure.

7 Experimental Results

Our experiment compares the original English queries to three retranslations generated by different versions of the transfer dictionary. The first version uses the dictionary constructed automatically by the process described in the previous section. Given that many dictionary entries have extraneous terms and sometimes the correct definition is missed completely, we decided to also build a clean version of the transfer dictionary manually for the query terms used in the experiment. Since the queries are short with some duplication, there are only about 300 unique terms to look up. We used the 3rd edition Robert and Collins French-English dictionary (not online) for this work. In hindsight, we realized that using different dictionaries for automatic vs. manual translation³ may cause some inconsistency in the results, largely because one dictionary may be more comprehensive than another, resulting in more definitions for some terms. The second query representation is generated from this manually constructed transfer dictionary.

Original English	Automatic word-based transfer dict.	Manual word-based transfer dict.	Manual multi-word transfer dict.
0.393	0.235	0.269	0.357

Table 2: Average Precision at 5, 10, 15, and 20 documents retrieved for the original English queries and translation via three different versions of the transfer dictionary.

During the manual construction process, we realized that the translation of multi-word noun phrases as an individual unit is particularly important. The automatically created transfer dictionary provided word-based translation only. Therefore, whenever we found a multi-word expression (MWE) in the bilingual dictionary which was also matched in the query, we added it to the transfer dictionary. This multi-word transfer dictionary serves as the basis for the third query representation. In summary, we generated four different experimental runs: (a) the original English queries and translated queries constructed from (b) the automatically generated word-based transfer dictionary, (c) the manually built word-based transfer dictionary, and (d) the manually built multi-word transfer dictionary. Experiments (c) and (d) are artificial, in the sense that this level of performance could not be obtained for a different set of queries without additional manual effort. They are included to help us understand the inherent limitations of this methodology by factoring out problems with the current implementation. The evaluation results averaged over the 50-query sample are presented in Table 2.

When we analyze the average performance figures using the Friedman Test [16], we find that queries (b) and (c) perform significantly worse than queries (a) and (d), but that the difference between translation by the multi-word transfer dictionary (d) and the original English queries

³We remind the reader that in this context *translation* means replacing a source language term with all entries from the transfer dictionary. No lexical ambiguity is resolved.

(a) is not statistically significant. While there is probably a real difference present which we were unable to detect due to the small size of the query sample, these results indicate that an IR system can perform almost as well across languages as it can in a monolingual setting, provided that a sufficiently accurate and comprehensive transfer dictionary is available. Correct translation of multi-word expressions makes the biggest difference in average performance.

There are several caveats which should be applied to these results. We had to manually construct the multi-word transfer dictionary in order to obtain the best results. Building a similar dictionary with full coverage of the language would be an immense task, therefore the final column of Table 2 should be interpreted as an optimal performance benchmark for transfer dictionary-based translation and represents a level of performance which is not currently attainable. Second, most modern IR systems have many additional features, such as automatic query expansion or phrase matching which improve performance. Our original English performance baseline has none of these features (i.e. MWE's are scored by individual term matching only). Were such techniques applied, the gap in performance between the original and translated queries could well increase. There are other ways that we might be able to improve the English baseline, such as adopting term-weighting strategies that are better designed for short queries.

7.1 Detailed Query Analysis

The average performance analysis only shows part of the picture. In order to get a different view of the results, we break down the performance by query in Table 3. We begin by removing the 8 queries whose original English version have an average precision of 0.0 from the sample (corresponding to no relevant documents ranked in the top 20). The goal of the MLIR system is to obtain performance equivalent to its monolingual counterpart. However, there is an important difference between systems that perform equally well at a high level and equally poorly (i.e. score zero according the evaluation measure). In the top half of the table, we arbitrarily define an absolute difference of average precision of more than 0.10 as important and partition the remaining queries according to whether their translated versions fall above or below this threshold. In the bottom half the table, we divide the queries that perform worse in translation according to whether they have an average precision greater than zero.

We notice a steady improvement in performance as we move from the automatically generated to manually generated dictionaries and add multi-word expressions. This difference is reflected primarily in queries moving from having no relevant documents ranked in the top 20 to at least some relevant documents scoring well, which is important. Recovering at least one relevant document is substantially better than finding none, because monolingual relevance feedback becomes a viable option. There are even some queries that perform much better in their translated versions.

In order to gain a better understanding of the problems associated with query translation,

Performance	Automatic word-based transfer dict.	Manual word-based transfer dict.	Manual multi-word transfer dict.
Tr > Orig	1	3	4
Tr \approx Orig	19	22	26
Tr < Orig	22	17	12
0.0 < Tr < Orig	10	9	9
Tr = 0.0	12	8	3

Table 3: Histogram comparing the performance of the translated (Tr) and original (Orig) English queries. Values given are the number of queries in each category.

we selected the 17 queries which did worse when translated using the manual word-based transfer dictionary and performed a detailed failure analysis. We found that 9 lost information as a result of the failure to translate multi-word expressions correctly, 8 had problems due to ambiguity in translation (i.e. extraneous definitions added to query), and 4 suffered from a loss in retranslation. Note that the total is greater than 17 because some queries suffered from more than one problem.

- Q53:** rachat financé par emprunt \implies leveraged buyout
- Q55:** délit d’initié \implies insider trading
- Q56:** taux d’emprunt préférentiels \implies prime lending rate
- Q58:** chemin de fer \implies railroad
- Q62:** coup d’état \implies coup d’etat
- Q64:** prise d’otage \implies hostage-taking
- Q65:** système de recherche documentaire \implies information retrieval system
- Q66:** langage naturel \implies natural language
- Q70:** mère porteuse \implies surrogate mother
- Q82:** génie génétique \implies genetic engineering
- Q88:** pétrole brut \implies crude oil
- Q90:** gaz naturel \implies natural gas
- Q96:** programme informatique \implies computer program

Table 4: A list of the important multi-word expressions in the query sample

Our experimental results demonstrate that recognizing and translating multi-word expressions is crucial to success in MLIR. This is in distinct contrast to monolingual IR, where identifying noun phrases or word pairs generally helps but does not produce dramatic gains in performance. The key difference is that the individual components of phrases often have very different meanings in translation, so the entire sense of the phrase is often lost. This is not always the case, but it happens often enough to make correct phrase translation the single most important factor in our multilingual experiments. Table 4 provides a list of the important MWE’s and their correct translations in our query sample. About half of these expressions lose vital semantic content when translated on a word by word basis.

Ambiguity in translation can also cause serious problems by adding noise to the query in the form of irrelevant translations. Here are some examples which illustrate the problem:

- machine \implies machine \implies machine engine
- amendment \implies amendement \implies amendment enrichment enriching agent
- measure \implies mesure \implies measure measurement moderation tempo
- failure \implies faillite \implies fail bankruptcy collapse failure
- military \implies militaire \implies military army serviceman
- affair \implies affaire \implies matter business affair case deal bargain transaction

In those case where rare and inappropriate definitions are added to the query (machine, amendment, measure), the ambiguity seriously hurts performance. In other cases (failure, military), the expanded terms all have similar meaning, and the results remain unchanged (since the query using failure is on bank failures). In the last example (affair - taken from *Iran-Contra affair*), the additional terms are valuable for query expansion, and cause a dramatic improvement in performance. Thesauri are legitimate tools for query expansion and this benefit can sometimes extend to the multilingual domain.

As mentioned previously, the simple concatenation of dictionary entries (giving all terms in the query equal weight) is a particularly naive approach to the problem of term weighting in query translation. However, it performs quite well given its simplicity. An alternative approach might be to adopt a probabilistic scheme that adjusts the weights according to the number of translations. For example, the three English translations of a French term would each get 1/3 of their normal term weight, corresponding to the uncertainty of the system regarding the correct definition. We tried this experiment, and found that the average precision dropped from 0.357 to 0.297 as a result of this measure. However, we should note that the 11-pt average precision remained constant and average precision of the reweighted queries was actually higher for 50% and greater recall.

How can we explain this result? A scan through the list of translations reveals that the occurrence of many translations is not necessarily a sign of uncertainty, as definitions are often close synonyms of one another. Full weighting of the dictionary expansion of *affair* above sharply improved performance. While words with only a single translation appear in general to be valuable, there is no pattern that generalizes to more translations. Most modern IR systems weight each term fully in query expansion, so our naive approach to term weighting (in the absence of more valuable information) seems reasonable.

Translation is not a isomorphic mapping. There are many humorous examples of the decay in meaning after repeated translations [1]. In our experiments, there is only one intermediate language, but this alone can lead to confusion. We describe problems of this kind as loss in retranslation. It is important to remember that there is an additional source of error in our experiments that comes from the fact that we translate the queries into French before they are translated back into English. A French user working directly with the MLIR system would not have this additional layer of indirection. Here are some examples of this process (braces link multiple translations of the same term).

- financial crunch \implies pertes financières \implies [loss ruin waste] financial
- proven \implies confirmé \implies confirmed
- demographic shifts \implies déplacements de population \implies population [movement displacement transfer travel trip]
- valued \implies représentant \implies represent stand portray depict representative

Many of these examples could be correctly resolved with a perfect transfer dictionary, but the variety and richness of language is such that a complete reference of this kind cannot be created.

7.2 Sample Query Profile

English: original intent or interpretation of amendments to the U.S. Constitution

French: l'intention première ou une interprétation d'un amendement de la constitution des USA

Term vector retranslation:

intention – intention benefit
 première – first initial bottom early front top leading basic primary original
 interprétation – interpretation
 amendement – amendment enrichment enriching agent
 constitution – formation settlement constitution
 USA – USA

Table 5: An example of the query translation process for short query Q76

In order to give a concrete picture of the general problems discussed in the previous section, we present a detailed profile of a single query. Table 5 shows the original text and the results of retranslation using the manually-generated version of the transfer dictionary for query Q76. This query is definitely not a triumph for our system, as the average precision score of 0.54 for the English text decays to 0.05 after retranslation. The course of this decay is measured in Table 6.

version	average precision	reasons for decay
orig Eng	0.54	
LR	0.34	intent \implies intention, U.S. \implies USA
TA 1	0.19	constitution, amendment
TA 2	0.10	original, intention
trans Eng	0.05	

Table 6: The decay in performance of query 76 from the original English (orig Eng) to the translated English (trans Eng) due to translation ambiguity (TA) and loss in retranslation (LR)

The first drop is caused by loss in retranslation. These differences are very subtle, but surprisingly important for IR performance, and demonstrate that search success is highly dependent on seemingly random choices in word selection. Intent and intention are close synonyms, and even native speakers of English might not distinguish between them in this context. These particular errors are mostly due to the experimental methodology⁴, but they are representative

⁴Many traditional stemming algorithms (such as Porter) would normalize these words to the same root, but inflectional morphology maintains the distinction as each word has a separate dictionary definition.

of the kinds of problems that a non-native speaker might have in finding the best vocabulary to describe a query topic. The two subsequent drops are the results of translation ambiguity, due to the fact that each word is replaced by all of its translations. Clearly, words like enrichment, settlement, and formation are unrelated to the primary topic of the query. If this ambiguity could somehow be resolved, we could expect some real improvements in performance.

Fr: amendement	Fr: constitution	χ^2 score
amendment	formation	4
	settlement	71
	constitution	11961
enrichment	formation	8
	settlement	7
	constitution	9

Table 7: Chi-square scores generated from a likelihood ratio test applied to contingency tables of document cooccurrences

One possible approach to resolving ambiguity would be to use the target language text to determine which translations tend to cooccur together. If translations of different query terms are used in the same context then this is evidence in favor of these particular translations. For example, it would not be too difficult to identify *amendement de la constitution* as a noun phrase and look at the document cooccurrence patterns among the translations of the component terms. The statistical significance of the cooccurrence pattern for the contingency table generated by each pair of translations is measured using the likelihood ratio test [11] and presented in Table 7. The higher the score, the less likely that these words would occur together by chance. From these results, the correct translation can easily be recognized. However, when we repeated the same experiment for *intention première* we found no such evident pattern. It would also be difficult to determine which pairs of words to compare on a general basis. A serious effort to resolve ambiguity would require a much more comprehensive approach, but this example suggests that the development of such a strategy might be feasible. We plan to attack this problem in our future research.

8 Future Extensions

There are two primary sources of error in our current implementation of a multilingual information retrieval system, missing translations of multi-word expressions and unresolved ambiguity in word-based translation. In addition, there will be some loss in retranslation due to the experimental design which cannot be avoided (i.e. the ambiguity introduced by the human translator). We find a substantial gap in performance between the original English queries and the translated queries which are generated from our noisy automatic word-based transfer dictionary. However, the dramatic improvements that result from manual corrections and additions to this dictionary

indicate that with work, one can expect the MLIR system of the future to approach the performance level of its monolingual counterpart. We hope to move in that direction with our future efforts.

To obtain multi-word expressions, one could simply attempt to gather together terminology lists from various specialized domains. However, such resources are precious and tend to be carefully guarded by their owners, making them expensive and/or not easily available. A natural alternative to the direct approach is to perform terminology extraction from corpora. There has been extensive research on the automatic recognition of terminology translations in parallel corpora [25, 9] and even some work on using non-parallel domain-specific corpora [20]. Information extracted by these techniques could be used to supplement the transfer dictionary in an MLIR system.

The error due to ambiguity could probably be reduced with proper term weighting strategies, although this is a difficult problem. The term alignment work of IBM [4] directly generates vectors of translation probabilities which could be incorporated into an MLIR system. This might help to reduce the importance of rare translations. However, these translation probabilities are likely to be highly domain-dependent so it is unclear how much they would help performance unless the training corpus is closely related to the collection used for retrieval. A multilingual collection, a fraction of which was available in parallel form, would be ideal for this kind of experiment. The vector space model is based on a term independence assumption, which is a questionable approximation of the true nature of language. This becomes a particularly cogent problem when one considers that the translation probabilities generated by IBM's word alignment follow this assumption. A very rare translation may always be correct in a particular context. Given the amount of data required to reliably estimate a word alignment model, the additional goal of estimating context based translation rules is probably still out of reach, as this step goes a long way towards solving the machine translation problem.

The term vector translation model tends to assume that retrieval strategies rely solely on the vector space model. In many ways, a weighted Boolean model might be more appropriate for MLIR. Linking term translations by disjunction would represent an elegant solution to the term weighting questions discussed in previous section. User derived conjunctions between query terms could serve as a natural filter for extraneous translations. Weighted boolean models deserve serious consideration in the context of multilingual information retrieval.

Direct translation filtering strategies applied the target language text collection represent a promising direction of research. One very simple method was presented in Table 7. More complex strategies could be devised that filter using context from the document collection. This approach has the nice advantage that the same collection is being used for both disambiguation and retrieval, so domain relevance of the filtering process is guaranteed. In this light, the translation disambiguation problem bears a strong resemblance to term disambiguation in a

monolingual setting. In fact, a number of researchers [8, 3] have used cross-language relationships to help with disambiguation. Given the limited success of term disambiguation as a tool for IR [23], there is some question about whether we can hope to gain any benefits out of translation filtering. Translation disambiguation may well work better as an interactive tool. If the user has some familiarity with the target language, he or she could be given a list of items and asked to select the intended definition, or perhaps eliminate the few that are clearly not relevant. This would be a lot cheaper and probably a lot more effective than automatic disambiguation.

Our experiments examine a very basic approach to multi-lingual information retrieval, simply replacing each term by the concatenation of its translations, as found in a transfer dictionary. This approach can work quite well if the dictionary is robust and comprehensive and translations for multi-word expressions are available. The current challenge in MLIR is to find ways to automatically extract the terminology lists and translation probabilities that are not available in the current generation of bilingual dictionaries. There is also a need to explore more structured query models of the translation process and determine whether or how much automatic disambiguation is desirable. The established body of literature on MLIR is scant indeed, so there are plenty of challenges ahead.

References

- [1] SPY magazine. page 48, March 1989.
- [2] J. Broglio, J.P. Callan, W.B. Croft, and D.W. Nachbar. Document retrieval and routing using the INQUERY system. In *Overview of the 3rd Text Retrieval Conference (TREC-3), NIST SP 500-225*, pages 29–38, 1995.
- [3] P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. Word-sense disambiguation using statistical methods. In *Proc. of the Association for Computational Linguistics (ACL)*, pages 169–176, 1991.
- [4] P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.
- [5] C. Buckley, G. Salton, J. Allen, and A. Singhal. Automatic query expansion using SMART: TREC-3. In *Overview of the 3rd Text Retrieval Conference (TREC-3), NIST SP 500-225*, pages 69–80, 1995.
- [6] Chris Buckley. Implementation of the smart information retrieval system. Technical Report 85-686, Cornell University, 1985. SMART is available for research use via anonymous FTP to ftp.cs.cornell.edu in the directory /pub/smart.
- [7] Jean-Pierre Chanod. Finite-state composition of french verb morphology. Technical Report MLTT-005, Rank Xerox Research Centre - Grenoble Laboratory, 1994.
- [8] I. Dagan, A.Itai, and U. Schwall. Two languages are more informative than one. In *Proc. of the Association for Computational Linguistics (ACL)*, pages 130–137, 1991.

- [9] Ido Dagan and Ken Church. Termight: Identifying and translating technical terminology. In *Proceedings of the ANLP*, pages 34–40, 1994.
- [10] Mark Davis and Ted Dunning. A TREC evaluation of query translation methods for multilingual text retrieval. In *The 4th Text Retrieval Conference (TREC-4)*, 1996. To appear.
- [11] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.
- [12] D.A. Evans, S.K. Handerson, I.A. Monarch, J. Pereiro, and W.R. Hersh. Mapping vocabularies using latent semantics. Technical Report CMU-LCL-91-1, Laboratory for Computational Linguistics, Carnegie Mellon University, 1991.
- [13] Gregory Grefenstette. Comparing two language identification schemes. In *Proc. of Analisi Statistica dei Dati Testuali (JADT)*, pages 263–268, 1995.
- [14] Donna Harman. Overview of the 3rd text retrieval conference (TREC-3). In *Overview of the 3rd Text Retrieval Conference (TREC-3), NIST SP 500-225*, pages 1–19, 1995.
- [15] M. Hearst, J. Pedersen, P. Pirolli, H. Schütze, G. Grefenstette, and D. Hull. Xerox site report: Four TREC-4 tracks. In *The 4th Text Retrieval Conference (TREC-4)*, 1996. To appear.
- [16] David Hull. Using statistical testing in the evaluation of retrieval performance. In *Proc. of the 16th ACM/SIGIR Conference*, pages 329–338, 1993.
- [17] David Hull. Stemming algorithms - a case study for detailed evaluation. *Journal of the American Society for Information Science*, 47(1):70–84, 1996.
- [18] Thomas K. Landauer and Michael L. Littman. Fully automatic cross-language document retrieval using latent semantic indexing. In *Proc. of the 6th Conference of UW Centre for the New OED and Text Research*, pages 31–38, 1990.
- [19] D.W. Oard, N. DeClaris, B.J. Dorr, and C. Faloutsos. On automatic filtering of multilingual texts. In *Proc. of the 1994 IEEE Conference on Systems, Man, and Cybernetics*, 1994.
- [20] Carol Peters and Eugenio Picchi. Capturing the comparable: a system for querying comparable text corpora. In *Proc. of Analisi Statistica dei Dati Testuali (JADT)*, pages 247–254, 1991.
- [21] Khaled Radwan. *Vers l'Accès Multilingue en Langage Naturel aux Bases de Données Textuelles*. PhD thesis, Université de Paris-Sud, Centre d'Orsay, 1994.
- [22] Gerard Salton. Automatic processing of foreign language documents. *Journal of the American Society for Information Science*, 21:187–194, 1970.
- [23] Mark Sanderson. Word sense disambiguation and information retrieval. In *Proc. of the 17th ACM/SIGIR Conference*, pages 142–150, 1994.
- [24] Penelope Sibun and A. Lawrence Spitz. Language determination: Natural language processing from scanned document images. In *Proceedings of the ANLP*, pages 15–21, 1994.
- [25] Pim van der Eijk. Automating the acquisition of bilingual terminology. In *Proceedings of the EACL*, pages 113–119, 1993.
- [26] E. Voorhees, N.K. Gupta, and B. Johnson-Laird. Learning collection fusion strategies. In *Proc. of the 18th ACM/SIGIR Conference*, pages 172–179, 1995.