Adaptive Filtering of Multilingual Document Streams

Douglas W. Oard

College of Library and Information Services, University of Maryland, College Park, MD 20742
Email: oard@glue.umd.edu Tel: 301-405-7590

Abstract

The increasingly ubiquitous global information structure makes it possible to examine high-volume text streams that contain documents written in a variety of languages. Present monolingual adaptive filtering techniques learn profiles which reflect user preferences and then apply those profiles to reduce the volume of new documents that must be examined by the user to manageable levels. This paper presents three techniques for extending adaptive monolingual text filtering techniques to manage multilingual document streams. Experimental results are given which demonstrate that dictionary-based and corpus-based techniques achieve similar performance in this application. This observation motivates our development of a translation technique designed specifically for vector space text representations which can in principle exploit both dictionary-based and corpus-based techniques. Results of initial experiments with this technique are given and the potential advantages of the new technique are discussed. The paper concludes with a discussion of future directions for adaptive multilingual text filtering.

Keywords

Cross-language text retrieval; information filtering; latent semantic indexing; machine translation.

1 Introduction

The explosive growth of the Internet and other sources of networked information have made automatic mediation of access to networked information sources an increasingly important problem. Much of this information is expressed as electronic text, and it is becoming practical to automatically convert some printed documents and recorded speech to into electronic text as well. Thus, automated systems capable of detecting useful documents are finding widespread application.

One important type of automated text detection system is what we call a text filtering system. As described by Belkin and Croft, information filtering systems seek to sift through large volumes of newly generated information, passing on to the user only those which might be useful [Belk92]. This is essentially the same concept that Luhn earlier called "Selective Dissemination of Information" (SDI) [Luhn58], but the term "information filtering" is now more commonly used when the information in question is arriving over a computer network. The vast majority of information filtering research has been focused on filtering electronic text, but interesting work has been done with music, home videos, and other media as well [Oard97b].

Many of the existing text filtering systems require that the user provide an explicit "profile" which specifies their information needs. What we call "adaptive" text filtering systems seek to minimize or eliminate this burden by learning the profile automatically. In many research systems, users are allowed to provide ratings for documents that they have examined. We have adopted this approach for our experiments because it allows for straightforward implementation and it suits our evaluation methodology well. In the future, adaptive text filtering systems will likely also exploit the sort of "over the shoulder" observations of user behavior investigated by Morita and Shinoda [Mori94]. Regardless of the approach chosen, present adaptive text filtering systems are most effective when used to satisfy relatively stable and specific information needs because a substantial quantity of consistent training data can be accumulated over time.

By "multilingual" text filtering systems, we mean systems which can select useful documents from document streams that may contain several languages (English, French, Chinese, ...). This formulation allows for the possibility that individual documents contain more than one language, a common occurrence in many applications. Multilingual text filtering systems can be useful even if the user is able to read only a single language. When sufficient resources are available to translate selected documents, for example, performing filtering before translation can be significantly more economical than performing translation before performing filtering. But even when translation is not available, there are circumstances in which multilingual text filtering could be useful to a monolingual user. A researcher, for example, might find a research paper published in an unfamiliar language useful if that paper contains references to works by the same author that are in the researcher's native language. But the most significant applications of multilingual text filtering will undoubtedly be those which involve multilingual users.

Text filtering systems for which users specify the profiles manually can easily accommodate multilingual filtering if the character set used for text representation is appropriate for the desired languages. All that is needed is either manual or semi-automatic facilities to translate the user-provided profiles into each language in which documents might be detected. This is the approach used by Paracel's Fast Data Finder system.[1]

Development of an effective adaptive multilingual text filtering system is considerably more challenging, however. The most straightforward approach, providing separate adaptive monolingual text filtering systems for each language, would only provide acceptable performance in languages for which an adequate quantity of training data could be observed or provided by the user. The techniques we have investigated, by contrast, are all capable of using a profile that was learned from material in any language (or that contain several languages in the same document) to select documents in any language.

In summary, applications for which adaptive multilingual text filtering systems are appropriate can be characterized by the following four features:

- Text extracted from the available information sources provides a basis for identification of potentially useful information.
- Users believe that automatic assistance with development of a suitable representation for their information needs would be helpful.
- The information need is fairly stable and specific, making it possible to acquire and exploit evidence about that need over time.
- The text may be in any of several languages, perhaps even including more than one language within a single document, and the users can make use of that text regardless of the language it is written in.

2 Background

We are not aware of any prior work on adaptive multilingual text filtering systems, but the closely related problem of cross-language text retrieval has an extensive research heritage [Oard97a]. In text retrieval the goal is to respond to an unforeseen information need (a "query") with information from a relatively static document collection. In the cross-language text retrieval problem the query need not be expressed in the same language as the documents.

The first practical approach to cross-language text retrieval required that the documents be manually indexed using a controlled vocabulary and that the user express the query using terms drawn from that same vocabulary. In such systems a multilingual thesaurus is used to relate the selected terms from each language to a common set of language-independent concept identifiers, and document selection was based on concept identifier matching. In the hands of an skilled user who is familiar with controlled vocabulary search techniques, such systems can be remarkably effective. Of particular note, if well designed, controlled vocabulary cross-language text retrieval systems can be just as effective as similar techniques would be in monolingual applications. Controlled vocabulary cross-language text retrieval systems are presently widely used in commercial and government applications for which

[1] Paracel Inc., 80 South Lake Avenue, Suite 650, Pasadena, CA 91101-2616

the number of concepts (and hence the size of the indexing vocabulary) is manageable. Unfortunately, the requirement to manually index the document collection makes controlled vocabulary text retrieval techniques unsuitable for large-volume applications in which the documents are generated from diverse sources that are not easily standardized.

This limitation has motivated the search for approaches which are amenable to less well structured situations. Two types of techniques have been investigated: dictionary-based approaches and corpus-based approaches. Dictionary-based approaches essentially seek to extend the fundamental idea of a multilingual thesaurus by using bilingual dictionaries to translate the query into every language in which a document might be found. Two factors limit the performance of this approach. The first is that many words do not have a unique translation, and sometimes the alternate translations have very different meanings. Monolingual text retrieval systems face similar challenges from polysemy (multiple meanings for a single word), but this translation ambiguity significantly exacerbates the problem. This problem is particularly severe in view of the observed tendency of untrained users to enter such short queries (often a single word) that it would not even be possible for a human to determine the intended meaning (and hence the proper query translation) from the available context.

The second problem with a dictionary-based approach is that the dictionary may lack some terms that are essential for a correct interpretation of the query. This may occur either because the query deals with a technical topic which is outside the scope of the dictionary or because the user has entered some form of abbreviation or slang which is not included in the dictionary. As dictionaries specifically designed for query translation are developed, the effect of this limitation may be reduced. But it is unlikely to be completely eliminated completely because language use is a creative activity, with new terms entering the lexicon all the time. There will naturally be a lag between the introduction of a term and its incorporation into a standard reference work such as a dictionary.

Corpus-based approaches seek to overcome these limitations by constructing query translation techniques which are appropriate for the way language is used in a specific application. Because it would be impractical to construct large tailored bilingual dictionaries manually, corpus-based approaches instead analyze large collections of existing text and automatically extract the information needed to construct these application-specific translation techniques. The collections which are analyzed may contain existing translations and the documents that were translated (a “parallel” collection), or they may be composed of documents on similar subjects which are written in different languages (a “comparable” collection).

Present corpus-based approaches are limited by two factors. The most significant limitation is that a parallel document collection which uses language in a manner similar to that found in the application may not be available in a suitable form. Techniques based on comparable document collections may eventually overcome this limitation, but research on the use of comparable document collections for text retrieval is presently at a very early stage [Picc96]. While a translation technique developed from a parallel document collection can be used for unrelated applications, significant reductions in retrieval effectiveness should be expected.

The other limitation of corpus-based techniques is that even when a suitable document collection is available, the methods presently used to extract the information on which the translation technique will be based introduce errors as well. Much of the initial research on corpus-based techniques has emphasized statistical analysis and made little use of linguistic theory. This approach has led to remarkable success. In machine translation, for example, statistical approaches have demonstrated performance equal to that achieved by linguistically motivated approaches, and they have done so with considerably less manual effort [Brow93]. But it also introduces errors that no human translator would make because the statistical approaches which have been applied are based on word cooccurrence and sometimes words which are not translations of each other exhibit the same patterns of cooccurrence as words which are translations of each other. There is some evidence that the incorporation of relatively simple linguistic information can significantly improve the performance of corpus-based techniques, and this appears to be a promising direction for future research.

Dictionary-based approaches and corpus-based approaches can both be applied to large unstructured document collections. One technique, Cross-Language Latent Semantic Indexing (CL-LSI), has even demonstrated cross-language text retrieval effectiveness that is on a par with the within-language performance of that same technique [Duma96]. This result is significant because an adaptive text filtering system based on Latent Semantic Indexing achieved a selection effectiveness nearly equal to that of the best participating systems at the third Text Retrieval Conference (TREC-3) [Duma95]. But the reported retrieval effectiveness results for CL-LSI were achieved with an experiment design that matched the retrieval application to the characteristics of the parallel document collection that was used to develop the translation technique.

No corpus-based system that we know of has yet demonstrated cross-language text retrieval effectiveness on a par with the within-language effectiveness of the same underlying retrieval techniques in the absence of a perfectly matched parallel document collection. We have thus chosen CL-LSI as the foundation for a corpus-based adaptive multilingual text filtering approach and sought to measure the effect of introducing a mismatch between the parallel document collection and the way language is used in the document stream which must be filtered.

In order to provide a basis for comparison, we have also experimentally determined the performance of a dictionary-based adaptive multilingual text filtering system. Our results show that in this application CL-LSI was able to achieve filtering effectiveness measures that are competitive with those achieved by a dictionary-based system for the same application. Since the types of errors made by a corpus-based system may differ significantly from those made by dictionary-based systems, we have developed a new corpus based technique which improves on CL-LSI, achieving similar performance when used by itself, but designed with an internal representation that is better suited to integration with existing bilingual dictionaries. We believe that this is a first step towards combining corpus-based and dictionary-based approaches to produce an adaptive multilingual text filtering that can achieve performance closer that achieved by monolingual system than would be possible using either technique alone.

3 Adaptive Multilingual Text Filtering

For consistency we have based all three of our approaches a technique developed by Dumais for adaptive monolingual text filtering in which Latent Semantic Indexing (LSI) is used to develop relatively short feature vectors that describe the relevant training documents, and the mean of the relevant documents' feature vectors is used as the profile [Duma95]. LSI feature vectors describing newly arrived documents are then used to rank order the newly arrived documents in order of decreasing similarity with the profile using the cosine similarity measure.

LSI feature vectors are constructed by counting the frequency with which each term occurs in a document and then using those values as input to a function which reduces the number of features by accounting for similarities in word usage. This function is automatically constructed using a static document collection which is representative of the documents which are expected to arrive. Initially, a matrix is formed in which the frequency with which a single term (word, phrase, etc.) occurs in a document is stored at the row representing the term and the column representing the document. These weights are then adjusted to estimate the importance of each term for retrieval by applying row and column operations that emphasize terms which occur frequently in an individual document but rarely in the representative collection. [2] A rank revealing matrix transformation (the singular value decomposition) is then used to construct a low-rank approximation to this matrix in which it has been observed by Deerwester, *et al.* that the vector components which remain in each column appear to represent the conceptual content of the documents fairly well while suppressing the adverse effects of variations in the author's choice of terms to represent those concepts in any particular document [Deer90]. This is accomplished by retaining (in our application) only the largest 200 singular values. A useful side effect of the singular value decomposition is that matrix formed from 200 retained singular values and their corresponding left singular vectors is a linear function which maps any column of the original matrix to the corresponding low-rank (200 element) representation that we call a LSI feature vector. Additional details of the technique can be found in [Berr96].

With this matrix it is possible to produce LSI feature vectors for previously unseen documents by simply forming a surrogate for the column that would have been constructed for original matrix if the document had been available when the singular value decomposition was computed and then performing a single matrix multiplication. This is important in text filtering applications, because even approximate incremental computation of the singular value decomposition is an expensive process. Dumais has shown that explicit positive feedback about some of the documents in a collection can be used to form a useful profile by simply forming the arithmetic mean of the feature vectors which represent those documents [Duma95]. This "LSI-mean" profile can then be used to rank order newly arrived documents so that those which are most similar to the documents which were used to build the profile will appear near the top of the list. This is done by creating LSI feature vectors for the newly arrived documents, computing the cosine of the angle between the profile and each document, and then sorting the cosines (and their associated documents) in decreasing

[2] These weights are computed using the "ltc" function in the SMART text retrieval system that is available at ftp://ftp.cs.cornell.edu/smart/

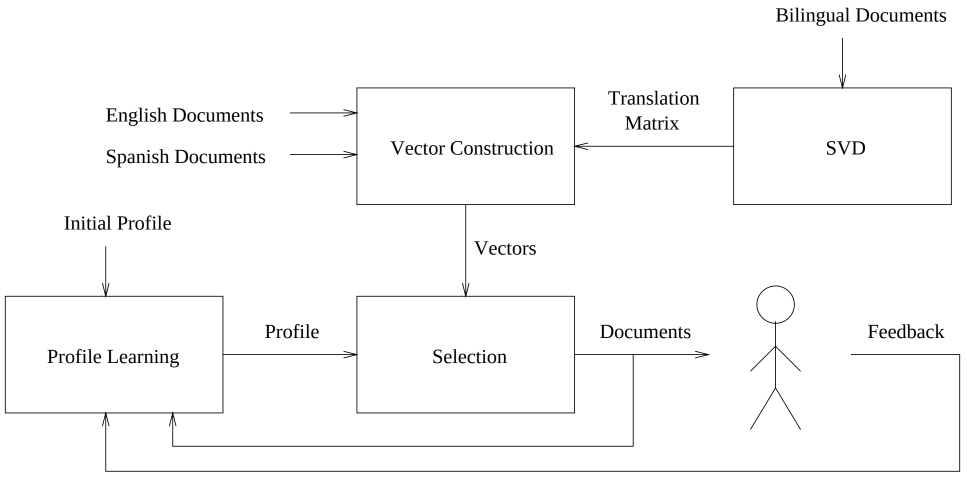

Figure 1: *Adaptive multilingual text filtering using Cross-Language LSI.*

order. Each cosine computation requires about 400 arithmetic operations (in general, about twice the number of retained singular values), and this is the slowest part of the LSI-mean filtering operation that must be performed for each newly arrived document. With the unoptimized code that we used in our experiments a SPARC 20 can compute about 50,000 cosines per hour for vectors of this length.

Extending LSI to CL-LSI is extremely straightforward. The only change which must be made is that the representative documents used to form the original matrix are formed by adjoining a translation of each document to the document itself. This requires that a parallel document collection be used for the initial step, so CL-LSI clearly satisfies our definition of a corpus-based technique. For consistency with our other approaches, we call this the "language training" step. The "profile training" and "filtering" steps that we described above can accept documents written in either a single language or in a combination of the two languages since no distinction is made between terms in the two languages when constructing the matrix of left singular vectors. Language-independent profile learning and text filtering is a natural consequence of combining CL-LSI with the LSI-mean adaptive text filtering technique in this way because LSI feature vectors are an inherently language-independent representation. Figure 1 illustrates adaptive multilingual text filtering using CL-LSI and the LSI-mean technique.

We used a more modular approach to construct our dictionary-based approach, but we were careful to retain as many features from our corpus-based design as possible in order to maximize the comparability of the results. In particular, we used a monolingual version of the LSI-mean adaptive filtering technique, training on only a single-language version of the same representative document collection. The language-independent nature of LSI feature vectors makes the query translation approach that is used in cross-language text retrieval impractical for text filtering applications based on the LSI-mean technique, so we chose instead to translate the documents themselves. In many real-world applications this approach would require that language identification be performed so that appropriate documents could be submitted for translation. This is not a significant obstacle, however, since language identification techniques with better than 95% accuracy are available [Gref95].

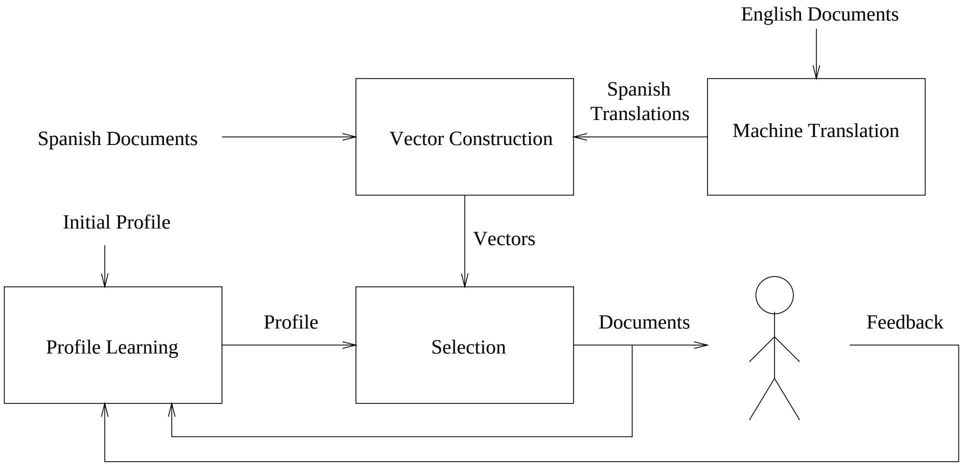

Figure 2: *Adaptive multilingual text filtering using Text Translation.*

In our experiments it was not necessary to perform language identification because each document used only a single language and the language of every document was known in advance. We used a fully-automatic machine translation system provided by the Logos corporation for our experiments.[3] This system contains a broad-coverage dictionary that can be augmented with application specific dictionaries, and software tools are provided to assist with their creation. In our experiments we used only the dictionaries delivered with release 7.0 of the Logos machine translation system. The Logos system also makes extensive use of linguistic information when performing translations. While some of this processing is unnecessary for text retrieval applications (e.g., word order choices), the effect of this information on word choice is significant. Logos generates only a single candidate translation regardless of the degree of ambiguity encoded in the dictionary, both syntactic information discovered during parsing and semantic information available in the dictionary (or "lexicon") is exploited when making these choices.

While such sophisticated processing is quite resource expensive, our experiment design required that we translate a fairly small number of documents. The within-language performance of the LSI-mean adaptive text filtering filtering technique has been well characterized, so we designed our experiments to measure the cross-language performance of our techniques. To do this we chose a monolingual profile training collection in one language and a separate monolingual evaluation collection in another language. A third, bilingual, language training collection was also used. With this design it was only necessary to translate documents in the monolingual profile training collection. Furthermore, since the LSI-mean technique exploits only positive feedback, it was only necessary to translate documents in that collection which were known to be relevant to the information need. For our profile training collection and the available relevance information this amounted to approximately 1000 documents. Because the Logos system we used was configured to translate documents from English into Spanish, we used an English language collection for profile training, a Spanish language collection for evaluation, and a bilingual English/Spanish collection for language training.

[3] Logos Corporation, 111 Howard Boulevard, Suite 214, Mount Arlington, NJ 07856 USA

The same collections and (where appropriate) relevance information were used in all of our experiments experiments.

In practice, the required translation effort would be much more extensive and it would need to be performed under tight time constraints. Thus, we believe that the approach we have chosen, which we call "Text Translation" (TT) is best viewed as an upper bound on the performance of practical dictionary-based approaches which exploit linguistic knowledge to select a single translation. It is possible that better integrated approaches such as the partial-translation approach used in the European Multilingual Information Retrieval project (EMIR) [Radw95] may eventually better this performance, but we believe our TT results presently provide a useful benchmark for the performance of a dictionary-based adaptive multilingual text filtering systems. Figure 2 illustrates adaptive multilingual text filtering using CL-LSI and the LSI-mean technique.

We developed our third technique, "Vector Translation" (VT), in order to overcome the computational bottleneck that is inherent in the TT approach. Our goal was to construct an effective corpus-based technique that is amenable to the introduction of linguistic knowledge as well. Unlike TT and CL-LSI, the VT technique is not inspired by an existing multilingual text retrieval technique. For that reason, we describe its motivation and operation in some detail.

Like TT, in VT every document is used to produce a Spanish vector. But in VT it is the document vector, rather than the document itself, that is translated. VT is essentially term-by-term translation applied to the vectors which represent documents (the columns in the term-document matrix that was described above). Since each element of a document vector is associated with a single term in a single language, term-by-term translation can be applied to vectors as easily as it can be applied to documents. Document vectors typically encode no term-order information, so deeper analysis is precluded by the representation. Term-by-term translation is quite fast, but as we observed above except in narrow domains where polysemy effects can be suppressed the resulting translations are generally not suitable for display to users.

Fortunately, the vector representation has two features which mitigate the adverse effects of this problem. The first is that it is not necessary to select a single translation target for each term, since the vector representation is based on the frequency with which a term occurs. For example, if there are two possible Spanish translations for some English term, it would be possible (although perhaps not wise) to simply divide the weight associated with the English term equally to produce a weight for each of the Spanish terms. The left side of Figure 3 illustrates this for two senses of the English word "bank," one of which (a financial institution) translates to "banco" and the other (a river bank) translates to "orilla."

The second helpful feature of vector representations is that a kind of "reverse polysemy" effect reduces the adverse impact of associating some of the weight with the wrong term in the other language. Consider the case in which two English terms both translate to the same Spanish term. The weights contributed by each English term can simply be added together to find the weight that should be assigned to the Spanish term. Word choice variation is a common stylistic device in many types of documents, typically introduced to avoid the monotony of repeated use of a single term. When different English terms are used for the same concept, it is likely that the intersection of their Spanish translations will be quite small,

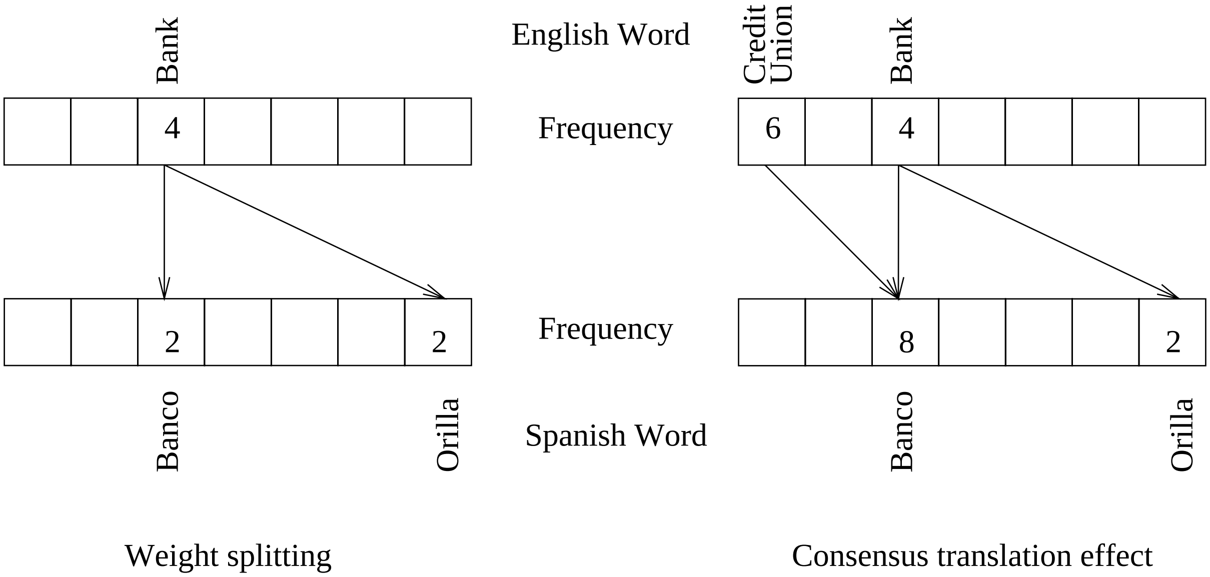

Figure 3: *Assignment of English term weights to Spanish terms.*

perhaps even a single word. Thus, term weight will tend to accumulate on the "consensus translation," and that consensus is likely to be correct. The right side of Figure 3 illustrates this consensus translation effect for the English terms "credit union" and "bank."

Figure 4 shows how this translation matrix is used for text filtering. Separate modules are used to construct vectors for English and Spanish documents because term lists and collection-wide statistics differ for the two languages. Vector translation is performed for vectors based originally on English documents so that all of the vectors passed to the translation module approximate those which would have been constructed if the original documents had been in Spanish. The remainder of the filtering process then proceeds as in the other two techniques. Vector Translation is considerably more efficient than Text Translation, requiring only a single matrix multiplication for each document that is not already in the preferred language. So once the translation matrix has been constructed, Vector Translation can be applied just as quickly as Cross-Language Latent Semantic Indexing

The translation matrix itself could be extracted from a bilingual dictionary if translation probabilities were associated with each word. Such information is lacking in presently available bilingual dictionaries, but in some cases the translations are listed in the order of their relative predominance. Even if this information could be interpreted in a useful way, however, this general order of predominance may not be very informative in specific applications for which only a few of the translations for any particular word would be appropriate. For this reason we have chosen to begin our work on vector translation using a corpus-based approach, and defer for future work the integration of information contained in bilingual dictionaries.

The corpus-based approach we use to build the translation matrix is based on the observed frequency of aligned terms that are extracted from a parallel bilingual document collection. Term alignment in bilingual document collections is a challenging problem which has been extensively studied as one important aspect of what is known as "corpus linguistics." Term

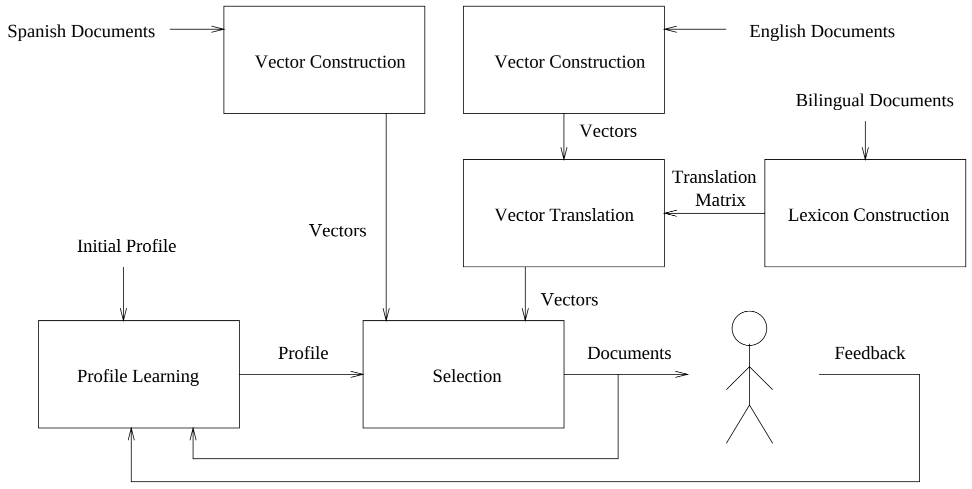

Figure 4: *Adaptive multilingual text filtering using Vector Translation.*

alignment typically proceeds in three stages:

1. Document alignment. Corresponding documents in each language are identified.

2. Sentence alignment. Sentences and other similar units in each text are identified and corresponding pairs of sentences are associated.

3. Term alignment. Corresponding terms (which may be words, word stems, morphological roots appearing in a dictionary, and/or multi-word phrases) in aligned sentence pairs are identified.

In our experiments we use a parallel bilingual collection of United Nations documents that has been aligned at the document level by the Linguistic Data Consortium at the University of Pennsylvania using document numbers assigned when each document was originally prepared.[4]. David Hull of the Rank Xerox Research Corporation then used a developmental version of morphological analysis software that is being developed at Rank Xerox to preprocess each document used in this experiment, converting each word to its morphological root and identifying commonly appearing phrases.

Shen and Dorr have developed a term alignment technique which Shen has used to produce aligned pairs from these preprocessed documents [Shen96]. They first use dynamic programming to optimize the alignment of sentences based on their length. Once the documents are aligned to at the sentence level, term-level alignment is performed. Their technique is based on the cooccurrence frequency of terms in aligned sentence pairs, with greater weight placed on cooccurrences that appear at similar locations within each sentence. For example, two words would be assigned a greater cooccurrence value if they both occur as the first word in a pair of aligned sentences than if one appears first and the other appears last. Every term pair with a cumulative cooccurrence value that exceeds a specified threshold is considered to be aligned. The number of such cooccurrences is then used to compute an empirical

[4] Information on the availability of the UN collection can be obtained from http://www.ldc.upenn.edu

distribution on the Spanish translation of every term appearing in the English versions of the United Nation documents, and those distributions are stored as a translation matrix.[5]

It is the use of a threshold on the correlation values which induces some measure of term alignment, and that is what distinguishes this approach from the vector translation technique of Davis and Dunning which was based solely on sentence alignment [Davi95]. A high threshold results in a sparse translation matrix and highly focused vector translations, a lower threshold produces more translation targets for each term and hence a somewhat more diffused translation in which the same amount of term weight is spread across a larger number of terms. One important goal of our experiments was to find a threshold value which produces enough diffusion to exploit the consensus translation effect, but which remains sufficiently focused to avoid introducing an overwhelming number of spurious translations.

4 Ideal Experiment Design

Evaluation of the three adaptive multilingual text filtering systems has been our greatest challenge. Experiments of the type we are conducting require a document collection for which relevance judgments are available, so it would be ideal if a large collection existed in which every document has versions in two languages and relevance judgments with respect to a number of standardized topics. The United Nations collection that we have used for language training satisfies the first part of this requirement, but there are no standard topics (and hence no relevance judgments) defined for that collection.

While we would ultimately like to provide users with systems which exploit new training data incrementally, when evaluating the filtering effectiveness of the algorithms themselves it suffices to introduce an artificial division between the construction of a profile and the use of that profile to rank order documents. This amounts to setting the “initial profile” in Figures 1, 2 and 4 to an empty vector, passing the known relevant documents in the profile training collection through the system to develop a profile, and then freezing this profile and using it to rank order the documents in the evaluation collection. If we had access to the sort of ideal test collection described above we could easily create language training, profile training, and evaluation collections by simply partitioning the collection three ways. Three partitions are needed to perform a fair experiment because it is not possible to use either the training collection or the evaluation collection for language training. In practical applications, cross-language techniques would simply not be needed if the documents in either of these partitions were already available in both languages. Of course, relevance judgments are not needed for the documents in the language training partition.

5 Use of Existing Document Collections

We are aware of no large bilingual collection for which relevance judgments are available, and large collections are needed for evaluation of adaptive text filtering systems. Furthermore, construction of topics and relevance judgments would have been well beyond our resources,

[5] The translation matrix is based on the number of identified alignments for each term pair, not on the strength of the observed correlations.

Source	English	Spanish	English Rel.	Spanish Rel.
1990-1993 UN Documents	X	X		
1990-1992 Wall St Journal	X		X	
1992 El Norte Newspaper		X		X

Table 1: *Evaluation using existing collections.*

so creation of such a test collection using the Linguistic Data Consortium's United Nations collection would have been impractical. Fortunately, over the course of five years the Text Retrieval Conferences have developed large monolingual collections with associated topics and relevance judgments, that can be used to augment the United Nations collection and create a useful (although somewhat awkward) test collection with the required characteristics.[6]

Because none of the three partitions must be both bilingual and scored, it is possible to use three separate collections to approximate the results that would be achieved using an ideal test collection. The collections we have chosen are shown in Table 1. TREC generates approximately 50 new topics each year, and the evaluation process produces a set of known relevant documents from parts of the TREC collection which vary from year to year. There are presently 300 TREC topics. One part of the TREC collection (a set of Wall Street Journal articles from what is known as TIPSTER disk 2) has been used each year, so there are relevance judgments for that "Wall Street Journal collection" for every available topic. These Wall Street Journal articles are mostly from 1990 and 1991, with a small number form 1992. We chose to restrict our language training collection to a similar portion of the United Nations collection in order to maximize the chances that terms associated with temporally specific events would be common to the two collections.

The most recent three TREC evaluations have included a "multilingual" evaluation in which monolingual text retrieval in languages other than English is evaluated. There are presently 50 topics available for a collection of 1992 Spanish language articles from the Mexican newspaper El Norte. We chose this collection because it was in Spanish (a language for which we had machine translation available) and because the time period overlapped with the temporal span of the other two collections we are using. Both the Wall Street Journal and the El Norte collections were preprocessed by Rank Xerox in the same manner as the United Nations collection.

The 50 Spanish (El Norte) topics were not a subset of the 300 English (Wall Street Journal) topics, but we were able to identify sufficient overlap between the formal TREC topic descriptions in several cases. We performed topic alignment manually, examining each of the 50 Spanish topics and then scanning the list of 300 English topics in order to identify possible matches. The detailed topic descriptions were then compared and a set of topic pairs which appeared to be closely aligned were selected. Table 2 shows the four Spanish topics for

[6] Translation of each document in the TREC collection into a second language would have been another alternative, but an enormous amount of effort would have been required to produce sufficiently good translations to give us confidence in our results.

ID	Abbreviated Spanish Topic	Rel	ID	Abbreviated English Topic	Rel
SP18	Foreign car makers in Mexico	183	290	Foreign car makers in the US	37
SP22	Mexican inflation	346	008	Economic projections	141
SP25	Mexican privatization programs	359	128	Privatization of state assets	97
SP47	Mexican cancer cause research	77	123	Carcinogen research & control	47

Table 2: *Closely related English and Spanish TREC topics.*

which we have found closely corresponding English topics.[7] Although the topic descriptions in each pair have some differences, there is sufficient apparent overlap to suggest that a minimal adjustment to the sets of relevant documents would result in comparable sets of documents in the two languages. In fact, our experimental results confirm that it is possible to use the relevance judgments without any adjustment when the goal is to compare different cross-language mapping techniques.

Two difficulties can arise when three existing collections are used in place of an ideal test collection. The first is that the subjects addressed by the UN, the Wall Street Journal and El Norte would be expected to differ significantly. We refer to this problem as a "domain shift," between the collections since it is caused by differences in the domain of discourse of the three collections. A potentially even more serious problem is that the Wall Street Journal and El Norte articles were judged against topics which are similar but not identical. We call this second problem "topic shift."

The domain shift between the UN documents and the El Norte articles is fairly easy to evaluate by running the Text Translation experiment a second time. In the second run we simply use the El Norte documents for language training instead of the Spanish UN documents. Since the rank-reducing mapping constructed using the left singular vectors will then be better suited to way language is used in the El Norte articles, the difference in filtering effectiveness will reveal the effect of the domain shift between the UN collection and the El Norte collection. We have not developed any similar technique to reveal the effect of the topic shift between either of those collections and the Wall Street Journal collection, however.

We can also estimate the severity of the topic shift effect, but the procedure is considerably more complex. Given the nature of the available test collections, the topic shift is an unavoidable consequence of introducing a second language. So the key to evaluating the impact of the topic shift is to compare the cross-language and within-language filtering effectiveness of the same adaptive text filtering technique. Again we base our approach on a modification to the standard Text Translation experiment. We first partition the El Norte collection into a training collection and an evaluation collection and then perform a monolingual evaluation. That removes the effect of the topic shift completely, but it also removes the effect of errors introduced by the machine translation step. So the second thing we must do is to measure the effect of these translation errors in isolation. We use a modification of the

[7] Some more weakly aligned topic pairs that might also be useful are identified in [Oard96].

standard CL-LSI experiment to do this. Recall that with CL-LSI, LSI feature vectors can be produced from either English or Spanish documents. If the Wall Street Journal articles are translated into Spanish before being used for CL-LSI profile training, the observed drop in filtering effectiveness would be entirely attributable to errors introduced by the machine translation step. Since these are exactly the same errors that affect the Text Translation experiment, this result will indicate how much of the difference between Text Translation and the monolingual run is attributable to translation errors.

Performing the complete topic shift experiment as we have described would require that both runs be evaluated using only half of the El Norte collection. We have not yet configured our software to accommodate this, so the results we report below were collected using the entire El Norte collection for both training and evaluation when attempting to measure the topic shift. Those results overstate the effect of the topic shift because they evaluate memory rather than prediction accuracy (in this one case we are performing a filtering effectiveness evaluation on the training set). But the results do provide an upper bound on the magnitude of the topic shift, and that upper bound proved to be adequate to recognize one case in which an extreme topic shift made an apparently well-aligned topic pair unusable. In every other case we have been careful to provide separate training and evaluation sets, so the limitations of our topic shift experiments to not extend to the other aspects of our evaluation.

6 Results

Since our primary objective is to determine the *relative* performance of three techniques, we have chosen an effectiveness measure focussed on the portion of the ranked lists produced by the three methods that are likely to show the greatest differences. We expect to find the largest absolute differences in the quality of the ranking near the top of the list. Since this is the portion of the list that is most likely to be examined by users in interactive applications, such an effectiveness measure also yields some insight into the absolute performance that users might observe in such circumstances. The effectiveness measure we have chosen to report is the precision (the fraction of the known relevant documents that have been found) at a fixed value of recall (0.1—the point at which 10% of the known relevant documents have been seen). In our experiments, this recall of 0.1 is achieved after 35, 36 or 8 relevant documents (for topics SP22, SP25 and SP47 respectively) have been found.

We used the SMART text retrieval system, modified locally to include the LSI-mean technique, for our experiments. We substituted the morphological roots provided by the Rank Xerox morphological tagger for SMART stemming because in future Vector Translation experiments we plan to exploit the resulting compatibility with the structure of existing bilingual dictionaries. For the runs reported in this section we built the documents vectors using individual words and did not take advantage of the phrases or part-of-speech tags added by Rank Xerox.[8] We have not yet completed the runs for topic 290 because of a delay in obtaining translated versions of the relevant documents that were processed under conditions identical to those used in the remainder of the runs.

Table 3 shows the precision at 0.1 recall for the three cross-language text filtering tech-

[8] A full description of the experimental conditions is presented in [Oard96].

	Technique			Bounds			
Topic Pair	CL-LSI	TT	VT	Chance	Lower	TT-max	Upper
SP22/008	0.17	0.17	0.12	0.01	0.06	0.28	0.64
SP25/128	0.08	0.10	0.10	0.01	0.03	0.10	0.18
SP47/123	0.09	0.05	0.01	0.00	0.00	0.16	0.47

Table 3: *Adaptive multilingual text filtering experiment results.*

niques, Cross-Language Latent Semantic Indexing (CL-LSI), Vector Translation (VT) and Text Translation (TT), and the remaining four columns provide performance bounds that are useful for comparison. The "chance" column provides a theoretic lower bound on performance, showing the precision that would be expected at any level of recall if documents were selected manually. The "lower" column shows an observed lower bound on the observed precision at 0.1 recall that results from the effect of proper names and other lexical items that are the same in both languages. This observed lower bound is computed by repeating the TT experiment, but omitting the machine translation step shown in Figure 2. The "TT-max" column shows the observed upper bound on the performance of Text Translation. In order to obtain comparable results, in our "TT" (and "VT") experiments we performed the SVD on the Spanish portion of the same bilingual document collection that was used to produce bilingual documents for the CL-LSI SVD. So the "TT" (and "VT") results include the same domain shift that that the "CL-LSI" experiment unavoidably incurs. The "TT-max" values show the precision at 0.1 recall that is obtained when TT SVD is constructed using documents from the evaluation domain (El Norte articles). Finally, the "upper" column shows the observed upper bound when El Norte documents are used for the SVD, profile training and evaluation. This represents the monolingual performance of the LSI-mean filtering technique on each Spanish topic.

The most significant observations that we can draw from these results is that adaptive multilingual text filtering appears to be practical and that the corpora we used are adequate to demonstrate this. Both corpus-based techniques and dictionary-based techniques have demonstrated much better performance than the lower bound runs on these topic pairs, despite the limitations in our ability to accurately measure absolute performance that results from the topic and domain shifts.

Another interesting observation is that the results without cross-language mapping exhibit a surprising amount of variation. We attribute this effect to the existence of words which are common to Spanish and English that are useful for recognizing documents that are relevant to some topics. This observation has led us to conclude that when the available corpora limit a cross-language text filtering or retrieval experiment to a small number of topics, a baseline run with no cross-language component is a simple way to gain some useful insight into the significance of the results.

The "TT-max" figures in Table 3 show the effect of the domain shift. In two cases out of three, the domain shift between the UN collection and the El Norte collection appears to be substantial but not overwhelming. The lack of a clear domain shift effect in the third

	Experiment Design		CL-LSI Profile Training	
Topic Pair	Multilingual	Monolingual	English WSJ	Translated WSJ
SP10/022	0.02	0.20	0.01	0.01
SP22/008	0.17	0.46	0.17	0.14
SP25/128	0.10	0.10	0.08	0.13
SP47/123	0.05	0.45	0.09	0.03

Table 4: *Preliminary topic shift results.*

case is at least partially explained by low upper bound on the effectiveness of the LSI-mean technique itself on topic SP25. This poor performance could result from a number of factors, but one possible explanation is that the relevant documents may have a multimodal distribution in the reduced rank LSI vector space.

Table 4 shows preliminary results which provide bounds on the magnitude of the topic shift effect. Results for a fourth topic pair which we tried, SP10/022, are shown as well in order to illustrate the topic shift effect clearly. Although it appeared from a manual inspection of the topic descriptions that topics SP10 and 022 were as similar as any of the other pairs we had chosen, these results clearly reveal that SP10/022 is not a useful topic pair. Again, the SP25/128 topic pair yields unusual and as yet unexplained results, actually increasing precision when translation errors are introduced. The remaining two topic pairs show relatively large topic shift effects (although these are only upper bounds) after considering the relatively small translation error effects.

7 Future Work

We have good reason to believe that the results reported here can be substantially improved upon by applying some of the lessons that are emerging from current research in cross-language text retrieval. Several teams working on that problem have reported dramatic improvements in performance when phrases are also used for building document vectors, presumably because translation ambiguity is limited when working with phrases rather than just isolated words(c.f., [Hull96, Radw95]). In some initial experiments with the SP22/008 topic pair we have increased the precision at 0.1 recall for Vector Translation from 0.12 to 0.20 by using phrases and also constraining the possible alignments using part-of-speech information.

We are particularly interested in further investigating the performance of Vector Translation because it offers the potential to integrate the corpus-based and dictionary-based approaches. We have observed that many of the candidate alignments produced by statistical techniques make no semantic sense, so we believe that it would be productive to use the set of known translations in an existing bilingual dictionary (and perhaps any information that is available about the relative predominance of those translations) as a stronger constraint on the alignment process. We call this idea "seeding" the distribution with the dictionary since the probability mass is constrained to accumulate only on the seeds that we have provided. While seeding the distribution would be expected to drive the translation matrix from one tailored to a domain towards one suitable for more general application, the improvement in alignment accuracy (and hence in the effectiveness of the VT technique) could be significant.

Another way of adding linguistic knowledge to the translation matrix would be to adjust the matrix by hand after it has been constructed. If performance analysis were to reveal unusually poor performance for the cross-language component of a Vector Translation system on a particular set of topics, the translation probabilities for terms associated with those topics could be examined by a domain expert who is fluent in both languages. If the values in the matrix appear to be counterintuitive, it would be possible to adjust them manually. Such a process is not likely not prove economically feasible for many applications, however, unless automated tools are developed to identify potentially poor translation probabilities and either suggest improvements or apply those improvements without human intervention.

Other techniques for term alignment have been proposed as well, so an experiment in which different term alignment techniques were applied might yield some interesting insights. Brown, *et al.* at IBM have collected term translation statistics using vastly more sophisticated techniques which directly handle word to phrase translation and take advantage of information encoded in word order [Brow93]. Their technique produces a translation matrix which is conditioned on sequences of English words rather than on a single word. Such a distribution is easily converted to one conditioned on the final term in the sequence by summing across the possible prefixes of that term, although it is not clear whether the result would be any more accurate than Shen and Dorr's simpler technique.

Another issue that we would like to investigate is automatic construction of comparable document collections. The documents arriving in a multilingual document stream have all of the characteristics of a comparable document collection except for the known alignment of documents on similar topics. If such alignments could be discovered automatically and then used to construct a translation matrix, periodic updates to the translation matrix could be performed using a collection with extremely similar characteristics to the documents which are likely to arrive in the near future. Sheridan and Ballerini have demonstrated a technique for accomplishing this when structured topic labels are associated with news articles arriving on a newswire [Sher96]. If dictionary-based techniques can be used to do this when such tags are lacking (and if corpus-based techniques which exploit comparable corpora prove practical), the resulting ease of updates could dramatically improve the performance and practicality of adaptive multilingual text filtering.

The other critical need is for better test collections. Although we have been able to estimate (or at least bound) the effects topic and domain shifts, it would clearly be better if what we have described as an ideal test collection were available with relevance judgments for documents in several languages with respect to an identical set of topics. There will be a new cross-language text retrieval "pre-track" in the 1997 TREC-6 evaluation, and in future years that track may produce such a collection. TREC provides an excellent venue for developing such a collection because documents which are good candidates for relevance assessment (those with a significant likelihood of being relevant) can be identified by a wide variety of systems that apply a broad array of techniques in both monolingual and multilingual settings.

This can significantly reduces the costs associated relevance assessment while minimizing the likelihood that systematic errors will result in a large set of relevant documents being missed.

8 Conclusions

We have described three approaches to adaptive multilingual text filtering and characterized the relative effectiveness of each. Of the three techniques, CL-LSI alone develops a language-independent representation for each document. Both Text Translation and the more practical Vector Translation require that the user (or system designer) choose a preferred language into which the documents or document vectors will be translated. Each technique introduces different sources of errors, so it is interesting to compare their relative performance. In Text Translation the principal sources of error are failure to recognize a Spanish word and incorrect resolution of translation ambiguity. In Cross-Language Latent Semantic Indexing the principal source of error is the inability to separate polysemous uses of a term in the training collection, which makes it difficult to conflate terms that have different sets of polysemous senses across languages into a single concept representation. In Vector Translation the largest source of error results from the use of inaccurate alignments when constructing the translation matrix.

Another interesting basis for comparison is the source of the information that is used to perform the translation. In CL-LSI this information is extracted automatically from documents which are aligned only at the document (or passage) level. Because the individual components of the CL-LSI translation matrix lack any understandable individual interpretation, human assistance with the construction of the CL-LSI translation matrix is precluded. The VT translation matrix, on the other hand, is constructed from bilingual document collections in which individual terms have been aligned. In addition to exploiting more fine-grained information, this approach gives the elements of the VT translation matrix a natural interpretation. Each is the probability that a specific English word will be translated to a specific Spanish word for documents in the domain of interest. Such information can be collected automatically from bilingual document collections, but it can also be constrained and corrected using additional information that is available from dictionaries. Since corpus-based and dictionary-based approaches introduce different types of errors, this joint construction approach could substantially improve filtering effectiveness.

We have also presented an evaluation methodology that can be used to gain insight into the performance of adaptive multilingual text filtering techniques using existing document collections. Since the domain shift effect we have described is inherent in any corpus-based technique (unless precisely the right language training collections are available), the ability to characterize the magnitude of the domain shift effect that we have demonstrated will be important whenever dictionary-based and corpus-based techniques are being compared. The topic shift effect, on the other hand, is strictly an artifact of our experiment design, an effect which can be eliminated by investing in the construction of test collections tailored for evaluating the performance of adaptive multilingual text filtering techniques.

While it would be unreasonable to try to draw broadly applicable conclusions from the three aligned topic pairs that we had available, our results have demonstrated that adaptive multilingual text filtering techniques are presently available which perform well enough for some applications. As additional lessons from the growing body of cross-language text retrieval research are combined with further advances in corpus linguistics and improved test collections, the range of applications to which these techniques can be productively applied can be expected to expand significantly. There clearly is a need for adaptive multilingual text filtering techniques, and our work has led us to conclude that there are a number of promising avenues to be explored which offer the potential to satisfy that demand.

Acknowledgments

The author would like to express his appreciation to Bonnie Dorr for the support and insight that led to this work, to Wade Shen for his assistance with term alignment, to David Hull of Rank Xerox Research Centre for his assistance with preprocessing the document collections, and to the Logos corporation for machine translation support. This work has been supported in part by DOD contract MDA9043C7217, ARPA and ONR contract N00014-92-J-1929, ARPA contract DACA76-92-C009 and NSF award IRI-9357731.

References

[Belk92] Nicholas J. Belkin and W. Bruce Croft. Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM*, 35(12):29–38, December 1992.

[Berr96] Micheal W. Berry, Susan T. Dumais, and Gavin W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595, December 1996. http://www.cs.utk.edu/∼lsi.

[Brow93] Peter F. Brown, Steven A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June 1993.

[Davi95] Mark Davis and Ted Dunning. A TREC evaluation of query translation methods for multi-lingual text retrieval. In D. K. Harman, editor, *The Fourth Text Retrieval Conference (TREC-4)*. NIST, November 1995. http://crl.nmsu.edu/users/madavis/Site/Book2/trec4.ps.

[Deer90] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, September 1990. http://superbook.bellcore.com/∼std/lsiPapers.html.

[Duma95] S. T. Dumais. Latent Semantic Indexing (LSI): TREC-3 report. In D. K. Harman, editor, *Overview of the Third Text REtrieval Conference*, pages 219–230. NIST, Department of Commerce, November 1995. http://www-nlpir.nist.gov/TREC/.

[Duma96] Susan T. Dumais, Thomas K. Landauer, and Michael L. Littman. Automatic cross-linguistic information retrieval using latent semantic indexing. In Gregory

Grefenstette, editor, *Working Notes of the Workshop on Cross-Linguistic Information Retrieval.* ACM SIGIR, August 1996. http://superbook.bellcore.com/∼std/lsiPapers.html.

[Gref95] Gregory Grefenstette. Comparing two language identification schemes. In *Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data*, December 1995. http://www.rxrc.xerox.com/research/mltt/Tools/guesser.html.

[Hull96] David A. Hull and Gregory Grefenstette. Experiments in multilingual information retrieval. In *ACM SIGIR-96 Conference Proceedings*, 1996. http://www.xerox.fr/grenoble/mltt/people/hull/papers/sigir96.ps.

[Luhn58] H. P. Luhn. A business intelligence system. *IBM Journal of Research and Development*, 2(4):314–319, October 1958.

[Mori94] Masahiro Morita and Yoichi Shinoda. Information filtering based on user behavior analysis and best match text retrieval. In W. Bruce Croft and C.J. van Rijsbergen, editors, *ACM SIGIR-94 Conference Proceedings*, pages 272–281. Springer-Verlag, July 1994. http://shinoda-www.jaist.ac.jp:8000/papers/1994/sigir-94.ps.

[Oard96] Douglas William Oard. *Adaptive Vector Space Text Filtering for Monolingual and Cross-Language Applications.* PhD thesis, University of Maryland, College Park, August 1996. http://www.glue.umd.edu/∼oard/research.html.

[Oard97a] Douglas W. Oard. Alternative approaches for cross-language text retrieval. In *AAAI Symposium on Cross-Language Text and Speech Retrieval.* American Association for Artificial Intelligence, March 1997. http://www.glue.umd.edu/∼oard/research.html.

[Oard97b] Douglas W. Oard. The state of the art in text filtering. *User Modeling and User-Adapted Interaction*, 1997. To appear.

[Picc96] Eugenio Picchi and Carol Peters. Cross language information retrieval: A system for comparable corpus querying. In *Workshop on Cross-Linguistic Information Retrieval*, pages 24–33. ACM SIGIR, August 1996.

[Radw95] Khaled Radwan and Christian Fluhr. Textual database lexicon used as a filter to resolve semantic ambiguity application on multilingual information retrieval. In *Fourth Annual Symposium on Document Analysis and Information Retrieval*, pages 121–136, April 1995.

[Shen96] Wade Shen and Bonnie Dorr. Alignment of bilingual terms using linguistically constrained feedback. Technical Report CS-TR-3666, University of Maryland, College Park, 1996. http://www.cs.umd.edu/∼swade/cs-tr.ps.gz.

[Sher96] Páraic Sheridan and Jean Paul Ballerini. Experiments in multilingual information retrieval using the SPIDER system. In *ACM SIGIR-96 Conference Proceedings*, August 1996. http://www-ir.inf.ethz.ch/Public-Web/sheridan/papers/SIGIR96.ps.