Translingual Information Retrieval: Learning from Bilingual Corpora

(AI Journal special issue: Best of IJCAI-97)

Yiming Yang, Jaime G. Carbonell, Ralf D. Brown, Robert E. Frederking
Language Technologies Institute, School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213 USA

{yiming,jgc,ralf,ref}@cs.cmu.edu

Abstract

Translingual information retrieval (TLIR) consists of providing a query in one language and searching document collections in one or more different languages. This paper introduces new TLIR methods and reports on comparative TLIR experiments with these new methods and with previously reported ones in a realistic setting. Methods fall into two categories: query translation and statistical-IR approaches establishing translingual associations. The results show that using bilingual corpora for automated extraction of term equivalences in context outperforms dictionary-based methods. Translingual versions of the Generalized Vector Space Model (GVSM) and Latent Semantic Indexing (LSI) also perform well, as does translingual pseudo relevance feedback (PRF) and Example-Based Term-in-context Translation (EBT). All showed relatively small performance loss between monolingual and translingual versions, ranging between 87% to 101% of monolingual IR performance. Query translation based on a general machine-readable bilingual dictionary – heretofore the most popular method – did not match the performance of other, more sophisticated methods. Also, the previous very high LSI results in the literature based on "mate-finding" were superseded by more realistic relevance-based evaluations; LSI performance proved comparable to that of other statistical corpus-based methods.

Keywords:

- Information retrieval
- Translingual or Cross-Language IR
- Statistical learning
- Corpus-based methods
- Generalized Vector Space Model

1 Introduction

Translingual information retrieval (TLIR) has begun to receive considerable attention in recent years with the increased accessibility of ever-more-diverse on-line international text collections, including centrally the World Wide Web. In spite of recent TLIR work [20, 12, 16, 23, 1, 24, 9], evaluations of different TLIR techniques on realistic retrieval tasks are rare. This paper reports our evaluation of the results of both newly developed TLIR techniques and re-implementations of previously reported techniques.

Translingual information retrieval (aka "multilingual" or "cross-lingual" IR) consists of providing a query in one language and searching document collections in one or more different languages. One can envision many ways to bridge the language barrier between query and collection. In this paper, we focus on query translation and methods based on automatically establishing translingual associations between queries and documents without the need to translate either.

2 MT-Based Methods for TLIR

The *machine translation methods* for TLIR require that either the query be translated into the target language, and the translation be used to search the target-language collection, or the collection be translated into the source language, and the original query be used to search. Let us consider the pros and cons of each approach:

- *Translation Accuracy* – Both human and machine translation [6, 19] require context to achieve accuracy. Translating isolated words in a query is unreliable, largely due to unresolved lexical ambiguity. Translating documents should yield greater accuracy.
- *Retrieval Accuracy* – Since documents contain far more information than queries, random translation errors should cause less degradation for the IR task in documents than in queries. Hence for both this reason and the above, document translation is preferable in principle. In fact, preliminary findings by Dumais *et al* [12] support this line of reasoning.
- *Practicality* – Many document collections are very large. Most are searched remotely. Some are proprietary; individual documents may be read or down-loaded, but the entire collection may not be copied or translated. Even if these problems were surmount-able, translating the collection may require inordinately long computation and massive storage, not to mention re-indexing the translated collection.

Because translating document collections is less practical, we report only on translating the query for TLIR. If the query were formulated as phrases, a full sentence, or a paragraph, we could apply MT systems far more reliably. However, experience shows that users typically prefer to give isolated words, or at best, short phrases to an IR system.[1] The question is how to best translate a set of isolated words, since full fledged MT is not applicable. We investigated three approaches:

1. *Dictionary-based Term Translation* – Look up each query term in a general-purpose bilingual dictionary, and use all its possible translations. This is a form of query expansion upon translation. Other forms of dictionary-based query translation methods have been reported before [8, 16, 1], and our results reported in section 5 are consistent with the dictionary-translation literature.
2. *Corpus-based Term Translation* – Use a sentence-aligned bilingual training corpus to find the terms that co-occur in context across languages, thus creating a corpus-based term-equivalence matrix. This is a new approach, where terms are translated based on co-occurrence frequency in the context(s) defined by the document collection. Its results reported in section 5 prove superior to the dictionary-based approach.

[1] LYCOS reports that their typical user queries for general web search are only one to three words long, although they are occasionally reformulated into longer queries.

3. *Corpus-based Term-to-Sentence* – Use the same type of aligned bilingual training corpus to extract full sentences that co-occur in the target language with query terms in the source language. This is a translingual adaptation of local context query enhancement. Term-to-sentence expansion may enhance recall, but at a cost in precision.

Since Term-to-Sentence performed very poorly in initial experiments, it will not be described further. Only general-purpose dictionary translation (called DICT or GLOSS below) and corpus-based term translation (called EBT below, for *Example-Based Term translation*) are further described.

All three MT-based methods used variations of the Pangloss Example-Based Machine Translation engine (PanEBMT) [3]. In general, EBMT systems [3, 18] use a large corpus of example pairs of previously translated sentences in order to find close matches and translations of words and phrases in context. The PanEBMT parallel corpus was derived primarily from the Spanish and English portions of the UN Multilingual Corpus [14], with an admixture of texts from the Pan-American Health Organization and ARPA MT evaluations. The total corpus contains some 685,000 sentence pairs – about 250 megabytes – after duplicated Spanish sentences have been removed. PanEBMT translates by finding the set of matches to a new text string (word, phrase or sentence) in the indexed bilingual corpus. Then, the translations corresponding to these matches are combined into candidate translations of the new text. Because queries contain more isolated terms than phrases or sentences, our query-translation experiment is unable to exploit the power of EBMT. Instead, we developed the term-in-corpus-context translation method.

2.1 Example-based Term Translation (EBT)

In order to create domain-specific or corpus-specific bilingual dictionaries automatically, we start from a large sentence-aligned bilingual corpus and generate a large thresholded term co-occurrence table[4]. The result was used as the dictionary for corpus-based (example-based) term substitution.

Co-occurrence dictionary generation is performed in two phases: First the co-occurrence matrix (indexed by source-language words on one axis and target-language words on the other) is generated. Each cell in the matrix represents the number of times the source-language word occurred in the same sentence pair as the target-language word. Given this matrix, we compute the conditional probability that if the term occurs in one language its counterpart (i.e. its candidate translation) also occurs in the other language within the same sentence pair, and vice-versa. If this probability is above a pre-set threshold *in both directions*, then the term translation is added into the dictionary. Should a term in one language co-occur with several terms in the other language with sufficient frequency to pass the conditional probability threshold, *all* are stored as candidate translations. The corpus-based term translation techniques are discussed in greater detail in [4, 3].

This method has the nice property that adjusting the filtering thresholds allows us to tune a trade-off: stricter thresholds prevent spurious translations, but significantly reduce the possible translations; more lenient thresholds produce better yields, at the cost of allowing more spurious translations.

A thesaurus which has been generated as described above can be further refined, increasing vocabulary size and reducing spurious translations, using an iterative process that applies a portion of the EBMT subsentential alignment algorithm to constrain which co-occurrences are added to the co-occurrence matrix. Although beneficial for dictionaries which are to be used directly for translation, refining the thesaurus in this manner proved to be slightly *detrimental* to performance on the translingual retrieval task for all but the smallest of training corpora – the refinement process removes much of the useful query expansion provided by collocations in the original statistically-derived dictionary.

Three separate training corpora were used to generate corpus-based thesauri: the full 250 megabytes of aligned Spanish-English text available to PanEBMT (consisting almost entirely of text from the UN Multilingual Corpus [14]), a 33-megabyte contiguous subset thereof, and a 12-megabyte corpus consisting of the training texts from our experimental corpus (described in Section 4) and the non-UN portions of the PanEBMT corpus. After tuning the thresholds, the best dictionaries extracted from the two larger corpora reached identical translingual performance; despite its much smaller size, the narrower focus of the 12-megabyte corpus permitted its best dictionary to outperform all others.

2.2 Dictionary-Based Term Translation (DICT)

For English-Spanish term translation we used a version of the machine-readable Collins Spanish-English Dictionary. Its performance should not be taken as the maximum achievable by this technique, since we had to invert the dictionary, which substantially reduces the vocabulary (from 51,500 to 27,200 words), in order to translate in the English-to-Spanish direction. It remains to be seen if a larger dictionary would improve performance, as rare words tend to have a rather small effect on overall performance in related tasks [30] (corroborated by our own experience with corpus-derived dictionaries on this task).

Due to the way in which our Spanish-English dictionary was built, inverting it provides the benefit of additional query expansion in some cases. This dictionary was originally built (for use by PanEBMT in word alignment within sentence pairs in the corpus) by looking up each unique word in the UN corpus in the Collins machine-readable dictionary. For words which were not found, a Spanish stemmer was applied and the resulting word root looked up. Thus, the Spanish-English dictionary contains the root forms of English words for many inflected Spanish words. After inverting the dictionary, the root forms of English words typically generate a multitude of inflected Spanish forms. The down side of the dictionary inversion is that most inflected English words are not found in the dictionary at all.

2.3 Manual Glossary (GLOSS)

In addition to the Spanish-English Collins dictionary, we also had hand-built glossaries from the PANGLOSS project available [13]. As with the Collins dictionary, these were created for Spanish-to-English translation, so we inverted the glossaries and extracted the single-word English entries. The extracted entries were then added to the inverted Collins dictionary to form the translation dictionary used as the basis of the GLOSS method.

Unlike Collins, the Pangloss glossaries provide not only properly inflected English translations, but also target the highest-frequency Spanish words specifically to ensure correct translations. As a result, although the combination of Collins and Pangloss glossaries increases the total vocabulary by less than 2000 words, the additional vocabulary consists primarily of inflected forms of the most frequent words.

3 IR-based Methods for TLIR

We extended three monolingual retrieval methods to translingual retrieval: Pseudo-Relevance Feedback (PRF)[5], the Generalized Vector Space Model (GVSM)[27], and the Latent Semantic Indexing (LSI) approach[10]. In each case, a translingual semantic correspondence between queries and documents is established based on a document-aligned bilingual training corpus, without requiring bilingual dictionaries or machine translation.

All these methods, PRF, GVSM and LSI, are variants of the vector space model (VSM) which was initially developed by Salton and is a fundamental paradigm in monolingual text retrieval[21]. To allow clear theoretical comparison of these IR-based methods, let us define the notation for VSM. Both queries and documents are represented using vectors of term weights in this model:

$$\vec{q} = (q_1, q_2, \ldots, q_m)^t$$
$$\vec{d} = (d_1, d_2, \ldots, d_m)^t$$
$$sim(\vec{q}, \vec{d}) = \cos(\vec{q}, \vec{d}) = \frac{\sum_{i=1}^{m} q_i d_i}{\sqrt{\sum_{i=1}^{m} q_i^2}\sqrt{\sum_{i=1}^{m} d_i^2}}$$

where $\vec{q}$ is the query vector, $\vec{d}$ is the vector of a document in a corpus, m is the number of unique terms (words or phrases) in the corpus after stop-word elimination and stemming, and q_i and d_i are the term weights in the query and the document, respectively. A term is typically weighted by $TF * IDF$, i.e., the product of within-document term frequency (TF) and the *Inverted Document Frequency* (IDF) of the term[21].

3.1 Pseudo-Relevance Feedback

Pseudo-relevance feedback (PRF) (aka "local feedback") is a variation of the classic relevance feedback (RF)[22]. Relevance feedback is a query expansion technique which adds terms in the *relevant* documents found in a initial retrieval to the query, and uses the expanded query for further retrieval. It typically improves performance in monolingual retrieval compared to not using it. PRF differs from the true relevance feedback by assuming the top-ranking documents retrieved are all relevant. It is simpler because no user relevance judgments are required; it is not always as effective as RF because the top-ranking documents often include some irrelevant documents that may be misleading. Both positive and negative evidence was found in empirical studies with respect to the effect of PRF on retrieval accuracy [15, 25]. As discussed in section 5, we also found PRF cuts both ways, depending somewhat on how the queries were formulated originally.

Our primary interest in PRF is to effectively cross the language barrier in translingual retrieval. Adapting PRF (and RF) to translingual retrieval is natural if a bilingual corpus is available[7, 1]. That is, once the top-ranking documents are retrieved for a query in the source language, their translation mates (the corresponding documents in the target language) can be used to form the query in the target language. Figure 1 illustrates the data flow for translingual RF and PRF. The retrieval criterion in PRF for monolingual retrieval is defined to be:

$$\vec{q'} = \vec{q} + \sum_i \{\vec{d_i} | \vec{d_i} \in \text{kNN}(\vec{q})\}$$

$$sim(\vec{q}, \vec{d}) = \cos(\vec{q'}, \vec{d})$$

where $\vec{q}$ is the original query, $\vec{q'}$ is the query after the expansion, $\text{kNN}(\vec{q})$ is the set of k Nearest Neighbors (most highly-ranked documents) retrieved using $\vec{q}$, and k is a pre-determined parameter whose value is empirically chosen.

Correspondingly, the retrieval criterion in PRF for translingual retrieval is defined to be:

$$\vec{q_t} = \sum_i \{\vec{g_i} | \vec{d_i} \in \text{kNN}(\vec{q_s})\}$$

$$sim(\vec{q_s}, \vec{d_t}) = \cos(\vec{q_t}, \vec{d_t})$$

where $\vec{q_s}$ is the query vector in the source language, $\vec{d_i}$ is the document vector in the source language and $\vec{g_i}$ is the document vector of its translation; $\vec{q_t}$ is the constructed query vector in the target language, and $\vec{d_t}$ is the target document in the search space. The length of each vector is m, the size of the term vocabulary after stemming and stop-word removal. Each element in the query and document vectors is weighted by $TF * IDF$.

3.2 Generalized Vector Space Model

A criticism of conventional VSM is that it uses terms as an orthogonal basis of the vector space, but terms are often not semantically independent. Wong *et al* proposed an alternative, namely the "generalized vector space model" (GVSM) [27], also referred to as "the dual space" [23] which uses word combinations (or individual documents) to form the basis instead of individual terms. Empirical studies showed somewhat better performance of GVSM over conventional VSM when using binary term weighting (a value of one for terms present, and zero for terms absent), while the comparison is inconclusive if more advanced term weighting was used in VSM[26]. Comparison of GVSM with PRF and LSI has not been

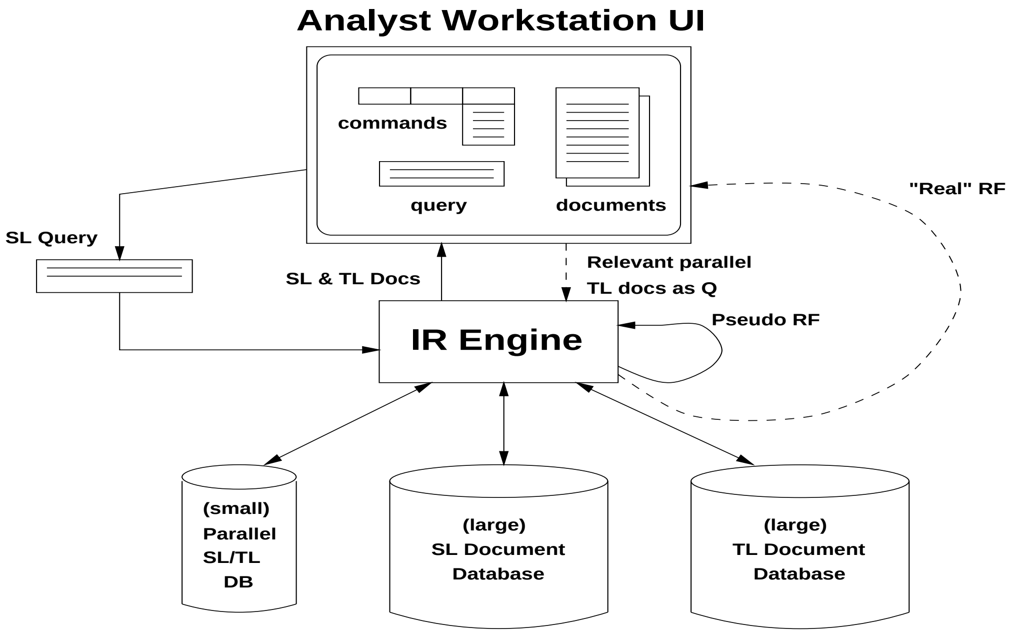

Figure 1: Data flow for translingual relevance feedback and pseudo-relevance feedback

carried out previously either in the monolingual retrieval literature or in the new TLIR literature. Our major focus here is a novel adaptation of GVSM to translingual retrieval. In order to thoroughly examine its properties, we will investigate its performance in both MLIR and TLIR, and compare GVSM with other methods including VSM, PRF and LSI.

The concept of the dual space can be explained using a term-document matrix. Given a document collection, one can represent this corpus using a matrix, $A_{m\times n}$, where the rows are unique terms in the document vocabulary, the columns are unique documents in the corpus, m is the vocabulary size, and n is the corpus size (number of unique documents). The elements in this matrix are within-document term weights, which can be binary-valued, or real-valued to combine within-document term frequency (TF) and corpus statistics, e.g., the Inverted Document Frequency or IDF of a term. One can view this matrix as a way to represent documents (the columns) using terms, and to represent terms (the rows) using documents. The former view corresponds to the conventional vector space model, and the latter view corresponds to GVSM in the dual space. The most interesting part of GVSM is the way a term is represented, i.e., each row vector of matrix A reflects the *pattern* of a term distributed over documents. Here lies the implicit assumption of this model: two words are semantically similar or highly relevant to each other if and only if they have a similar distribution over documents. This assumption is of course arguable given that "document" is often an arbitrary choice (it could be an abstract, an paragraph, a full article, a chapter, or a book). However, it provides a way to use corpus statistics to measure the closeness between terms, which is not directly offered by the conventional VSM.

It should be clarified that in the original GVSM model by Wong et al., unique word combinations are used as orthogonal dimensions, which is not equivalent to using unique documents. If two documents contain exactly the same set of unique words (although the term frequencies in these documents may be different), then these documents will have the same vector representation in the original GVSM. In this study, we made a simplification by directly using documents as dimensions in the dual space, assuming that different documents very rarely share identical vocabulary, and even if such cases do happen, their effect on the representation power of this model and its retrieval effectiveness may be negligible. This does not exclude possible applications that might benefit from using the original version, of course. We use GVSM to refer to our simplified version, with the assumption that our conclusions would apply to the original version of GVSM as well.

The monolingual version of this method (ML-GVSM) consists of query transformation, document transformation, and similarity comparison between the transformed query and documents. The retrieval criterion is defined to be:

$$sim(\vec{q}, \vec{d}) = \cos(A^t \vec{q}, A^t \vec{d}).$$

The query transformation, $\vec{q'} = A^t \vec{q}$, is equivalent to weighting the *distribution pattern* of each term (the row vector in A) using its weight in the original query, and summing up the weighted patterns to obtain a new representation of the query. The document transformation is similar: $\vec{d'} = A^t \vec{d}$ weights and sums up the the distribution patterns for the terms contained in the documents. The resulting vectors, $\vec{q'}$ and $\vec{d'}$, have n dimensions, corresponding to the n documents in matrix A. The document collection used in matrix A is usually called the training set, and the transformation of the query or a document is called *the fold-in* process. In general, the document to be transformed is not a member of the training set.

Our novel extension of the monolingual GVSM for translingual retrieval uses a bilingual corpus for training. Let us define two matrices, A and B, where A is a term-document matrix for the training documents in the source language (also the language of the queries), B is a term-document matrix for the training documents in the target language, and the corresponding columns of A and B are the matching pairs of documents in the bilingual corpus. These matrices are illustrated below (we use binary-valued elements here for simplicity):

		D_1	D_2	D_3	D_4	D_5	D_6	D_7	$\cdots$	D_{n-1}	D_n
	cat	0	1	0	1	1	0	0	$\cdots$	0	0
	dog	0	1	1	1	0	1	0	$\cdots$	0	0
A	*enter*	0	0	0	0	0	1	1	$\cdots$	1	0
	$\vdots$	0	0	0	0	0	0	0	$\cdots$	0	0
	lock	1	0	0	1	0	0	0	$\cdots$	1	0
	cerrar	1	0	0	1	0	0	1	$\cdots$	1	0
	finca	0	0	0	0	0	1	1	$\cdots$	0	0
B	*llave*	1	0	0	1	1	0	0	$\cdots$	1	0
	$\vdots$	0	0	0	0	0	0	0	$\cdots$	0	0
	perro	0	1	1	1	0	1	0	$\cdots$	0	0

where D_i is the ith pair of corresponding English and Spanish documents. A given source word, such as *dog*, is represented by its distribution in the *source*-language document set, while a given target word, such as *perro*, is represented by its distribution in the *target*-language document set. Words that are translations of each other often exhibit identical or very similar rows, as do *dog* and *perro*. Not all words have a one-to-one translation, and not all corresponding words have exactly the same occurrence pattern. For instance, the verb *to lock* is typically translated as *cerrar con llave*, accounting for their very similar occurrence patterns.

We use A for query transformation and B target-language document transformation. The retrieval criterion is defined to be:

$$sim(\vec{q}, \vec{d}) = \cos(A^t \vec{q}, B^t \vec{d})$$

Since matrices A and B share the same dual space, the transformations $A^t \vec{q}$ and $B^t \vec{d}$ give the query and the document a common basis (the distribution patterns of terms over documents) on which they can be compared. This is how the translingual correspondence is established.

The computation in GVSM consists of the transformation ($A^t \vec{q}$ and $B^t \vec{d}$) and the cosine computation. The time complexity of the first part is similar to the computation in VSM. It is proportional to the number of non-zero elements in a query or document vector, i.e., $O(kn)$, where k is the average number of unique terms per query or document, and n is the number of document pairs in the bilingual training corpus. The time and memory complexity in the second part, is $O(n)$ per document, or $O(nl)$ for a test corpus of l documents. This can be expensive for very large applications. Fortunately, we found it possible to significantly reduce this complexity by aggressively removing non-influential elements from the transformed document vectors without sacrificing retrieval performance, as shown in our previous work[28] and in the empirical results of this study (Section 5.4).

3.3 Latent Semantic Indexing

Latent Semantic Indexing[10] (LSI) is a one-step extension of GVSM. The claim is that neither terms nor documents are the optimal choice for the orthogonal basis of a semantic space, and that a reduced vector space consisting of the most meaningful linear combinations of the original dimensions would be a better representative basis for the content of documents.

In monolingual retrieval, LSI uses the term-document matrix (A) for training, the same as in GVSM. It computes the orthogonal dimensions ("the latent semantic structures") in matrix A, and selects the largest principal dimensions as the new basis for a reduced vector space. The monolingual LSI retrieval criterion is defined to be:

$$A = U\Sigma V^t$$

$$sim(\vec{q}, \vec{d}) = \cos(U^t\vec{q}, U^t\vec{d})$$

where matrices U and V contain a set of p orthogonal singular vectors each (one for the representation of terms, and another for the representation of documents). Matrix Σ is p-diagonal, containing the singular values indicating the importance of the corresponding singular vectors in matrices U and V. Matrix U can be viewed as a reduced version of matrix A in the sense that both A and U use their row vectors to represent terms, but the term vectors in U are much shorter than the term vectors in A. The dimensions in U are linear combinations of documents, while the dimensions in A are individual documents.

The translingual LSI model [12] is similar to the model for monolingual LSI, except that a bilingual document corpus is needed for training instead of a monolingual corpus. Let $\vec{q}$ be a query in the source language, $\vec{d}$ be a document in the target language, and $\begin{bmatrix} A \\ B \end{bmatrix}$ be the matrix of bilingual document pairs where A and B are the same as defined in GVSM. Then the translingual LSI retrieval criterion is defined to be:

$$\begin{bmatrix} A \\ B \end{bmatrix} = U_2\Sigma_2 V_2^t$$

$$sim(\vec{q}, \vec{d}) = \cos(U_2^t\vec{q}, U_2^t\vec{d})$$

where U_2, V_2 and Σ_2 are the matrices computed using the singular value decomposition of the bilingual input matrix .

LSI has a quadratic time complexity of $O(n'p)$ where $n' = \max\{m, n\}$ is the larger number between the size (m) of the joint vocabulary of both languages and the number (n) of document pairs in the bilingual training corpus; and p is the number of orthogonal dimensions (singular vectors) computed in the singular value decomposition. Thus, the scalability of this method to a large corpus would be much more limited than the VSM or GVSM approach if a large number of singular vectors is necessary for good retrieval performance.

3.4 The Scientific Challenge

The similarities and differences between the three models mentioned above can be seen in their retrieval criteria:

$$VSM : sim(\vec{q}, \vec{d}) = \cos(\vec{q}, \vec{d})$$

$$GVSM : sim(\vec{q}, \vec{d}) = \cos(A^t\vec{q}, B^t\vec{d})$$

$$LSI : sim(\vec{q}, \vec{d}) = \cos(U_2^t\vec{q}, U_2^t\vec{d})$$

In theory, the fundamental difference between these methods is the choice of the basis for the similarity comparison between queries and documents. VSM (including PRF) assumes semantic independence of terms in its basis. GVSM uses documents instead, assuming documents are semantically independent. LSI computes the orthogonal dimensions in a training corpus, and chooses the principal dimensions as the basis of a reduced vector space. GVSM and LSI are close variants in the sense that both exploit the dual space. The only difference between these two is the choice of the original dimensions (terms in documents) or the reduced dimensions (the orthogonal singular vectors) as the basis for the vector space. Which model best represents the semantic space of documents and queries is a scientifically challenging question.

Given these methods, empirical validation is important. For monolingual retrieval, performance improvement of GVSM over VSM was observed on small collections [27]; sometimes, improvement of LSI over VSM was observed, but not always [10]. Until the work described in this paper and its previous version[7], a comparison between GVSM and LSI in either monolingual or translingual retrieval has not been made.

4 Corpus and Query Preparations

In order to conduct an empirical evaluation, our first task was to prepare a bilingual corpus for translingual experimentation. The large UN Multilingual Corpus (about 500 megabytes of data per language) [14] from the Linguistic Data Consortium was available to us, but the original UN corpus is a heterogeneous mixture of many types of documents. Using formatting codes and alignment methods, we then extracted and segmented a subset of the data, consisting of 2255 document pairs pertaining to UNICEF reports and deliberations. We randomly selected 1134 document pairs for training, and set the remaining 1121 pairs aside for testing. Of these 550 documents were used as the validation set to test each method with different parameter settings, and 571 were used for the final blind testing reported in our results below. Altogether, the training and test sets in both languages consist of almost 2 million words of text, equivalent to about 22 megabytes of data. Each document has approximately 6 paragraphs; there are about 5 sentences per paragraph on average.

The second task was to develop queries and human relevance judgments for the evaluation of retrieval methods. We created 30 queries in English, germane to the UNICEF sub-collection. We then contracted externally for human relevance judgements on the cross product of the 30 queries and 1121 test documents (33,630 judgments in all), which are used as the gold standard for evaluation. This test set was further divided into a 550-document validation set (for experiments to optimize parameters for each method) and a 571-document blind test set for the final reported results. The query length varies from 6 to 36 words, with an average of 14 words per query. The number of relevant documents for a query varies from zero to 70 (one of these queries has no relevant documents), with an average of 16. For the blind test set of 571 documents, seven queries have no relevant documents.

To explore the effect of the granularity of corpus alignment on the performance of retrieval methods which are trained using the bilingual corpus, we further developed two additional versions of the training corpus: a paragraph-level alignment (7227 paragraph pairs) and a sentence-level alignment (21,591 sentence pairs)[2].

To further investigate the effectiveness of our methods, we also used a classic document collection, MEDLARS, commonly used in monolingual retrieval evaluation prior to the Text Retrieval Conference (TREC)[11]. MEDLARS contains 1033 documents and 30 queries, and provides human relevance judgments. The query lengths range from 2 words to 62 words, with an average 20 words per query. The number of relevant documents given a query is between 9 and 39, with an average of 23.

[2] The aligned parallel corpus, including training and testing partitions are made available by Carnegie Mellon University (CMU) to LDC members – email: yiming@cs.cmu.edu.

5 Empirical Evaluation

We conducted a comparative evaluation of our translingual IR methods on the UNICEF test set, including DICT, EBT, PRF, GVSM and LSI. The experiments were carried out as follows:

First, we trained each corpus-based method that requires off-line training, in order to find translingual equivalences using paired documents, without queries; hence no relevance judgements were required for training. Second, we tuned parameters for both the *monolingual* and *translingual* versions of our methods on the validation test set. Third, we measured the effectiveness of each method on the blind test set. Fourth, we evaluated the results by comparing the retrieval degradation when moving from monolingual to translingual IR for all the methods on the blind-test set. Fifth, we repeated the above experiments using paragraph-level alignment and sentence-level alignment instead of document-level alignment, and compared the behavior of the TLIR methods. Sixth, we tested LSI and GVSM in *mate finding* (described later), to see how difficult or how easy a task it is compared to realistic retrieval. Seventh, we tested our MLIR methods on the MEDLARS corpus and contrasted them with our observations on UNICEF, in order to compare the relative difficulty of retrieval in different document collections, with different queries.

5.1 Off-line training

The training phase differs by method. In fact, only EBT and LSI require off-line training. In EBT (example-based term substitution), training means extracting highly correlated English-Spanish term pairs in context from the bilingual corpus; these pairs are used later for term substitution in each query for translingual retrieval. In LSI, training means finding the principal orthogonal vectors by applying singular value decomposition (SVD) to matrix A; these orthogonal vectors are used for query and document transformation in the document indexing phase (called the *fold-in* phase). GVSM does not have a true training phase like LSI does because it directly uses the column vectors in the term-document matrix A as the basis for the dual space. PRF also does not have off-line training because its query formulation is based on on-line retrieval of documents given a query. DICT (query translation using a machine-readable dictionary) does not make any use of the bilingual corpus and therefore does not require a training phase (although there is still preparation involved prior to use).

5.2 MLIR experiments and parameter optimization

We measured retrieval performance using the conventional 11-point average precision metric. For a retrieval system that produces a ranked list of documents given a query, the performance is typically measured using the average precision over different recall levels. Recall is the ratio of retrieved relevant documents over the total relevant documents in the collection; precision is the ratio of retrieved relevant documents over the total retrieved documents. The 11-point average precision is the *interpolated* average of precision values when thresholding at recall levels of 0%, 10%, ..., 100%. For further details of the interpolated averaging method, refer to [21]. For brevity we refer simply to "average precision" or AVGP.

In the monolingual retrieval experiments, we optimized each method with respect to its performance on the UNICEF corpus using the human relevance judgements on the 30 queries and the 550 validation documents. Optimizations include:

- Determining the best term weighting scheme (using a combination of TF and IDF, for example) for cosine-similarity scoring. This optimization is applied in all the methods. We found the SMART ntc.ntc term weighting optimal for all the methods on UNICEF, meaning that both document and query terms are weighted linearly by TF*IDF with cosine normalization.
- Rank-based thresholding on the retrieved documents in pseudo-relevance feedback, i.e., labeling the top k documents to use as relevant for query expansion. We found k=10 optimal for UNICEF. The second parameter, SP (for *sparsification*), optimized in PRF is the number of most influential terms retained in the query after PRF expansion. We found SP=70 optimal for PRF.

Method	ML.	TL	TL/ML	Corpus align
Dict♯	(0.4884)	0.3901	80%	N/A
GLOSS♯	(0.4884)	0.4064	83%	N/A
EBT♯	(0.4884)	0.4918	101%	sentence
PRF (SP=70, K=10)	0.4255	0.4203	99%	paragraph
GVSM (SP=200)	0.5035	0.4585	91%	paragraph
LSI (SV=200)	0.4884	0.4234	87%	document

♯ The result of SMART.basic is used as the performance baseline.

Table 1: CMU results in monolingual and translingual retrieval on UNICEF

Site	Method	ML	TL	TL/ML
UMASS	correlated phrases	.20	.1358	68%
ETH	Similarity thesaurus	.527	.212-.278	40-53%
XEROX	Dict	.393	.235	60%
NMSU	Dict	?	?	73.5%

Table 2: Published results in ML and TL retrieval on other corpora

- Determining the GVSM sparsification (SP) parameter corresponding to the number of most influential terms per vector retained after query or document transformation. All other terms are zeroed out. We found SP=200 optimal for GVSM.
- Determining the optimal number of singular values (SVs) to compute in LSI, meaning the number of most influential orthogonal dimensions used. We found that SV=200 reaches a performance plateau for LSI in UNICEF, which is exceeded only as SV approximates 1000. However, at that level the number of retained dimensions is approximately equal to the number of original dimensions, and LSI provides no dimensionality-reduction benefit. Moreover, the SVD step is computationally very expensive for large SV, and convergence of the sparse SVD algorithm is particularly slow as SV approximates the original dimensionality. Therefore, since the marginal improvement of SV=1000 over SV=200 comes at a very steep price, we select 200 as the more reasonable SV value.

The parameter values for each method which produced optimal performance in monolingual retrieval were also found optimal or nearly optimal in our translingual experiments (Section5.4).

We implemented all the monolingual and translingual methods using components of the publicly-available SMART retrieval engine [21], including indexing, stemming, TF*IDF-based word weighting and stop-word elimination in both languages[3]. This common infrastructure enabled us to factor out extraneous variables from our experiments. For the monolingual VSM baseline, we ran SMART without relevance feedback (SMART.basic).

5.3 Primary results

Figure 2 presents the recall/precision curves for the MLIR methods on UNICEF, and Figure 3 the recall/precision curves for TLIR methods. The 11-point average precision values of these methods are summarized in in Table 1. We also conducted a different evaluation without using human relevance judgments of the same methods, instead relying on the degree of overlap between documents retrieved monolingually and their translation-mates retrieved translingually. Such an evaluation, as reported in our previous paper[7], may not be as informative, but is helpful when human relevance judgements are not available.

[3] We expanded the SMART Spanish stop word list so that its coverage is equivalent to the English one, resulting in a somewhat longer list because of irregular inflections for Spanish auxiliary verbs.

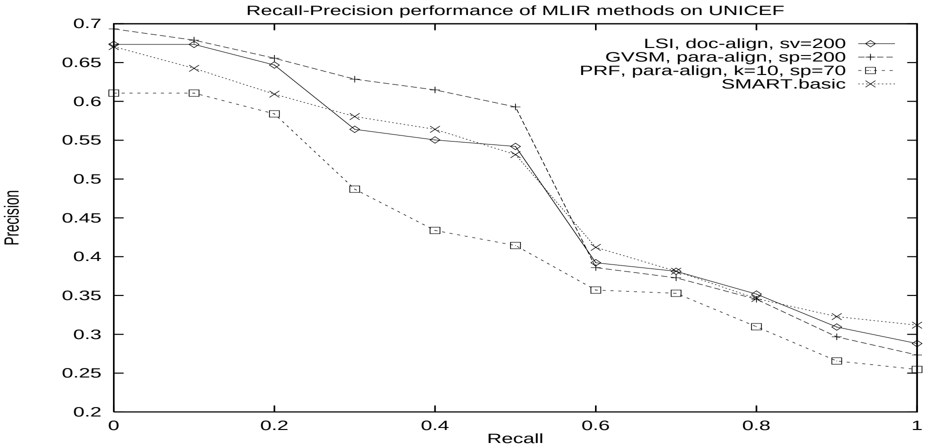

Figure 2: Recall-Precision performance of MLIR methods on UNICEF

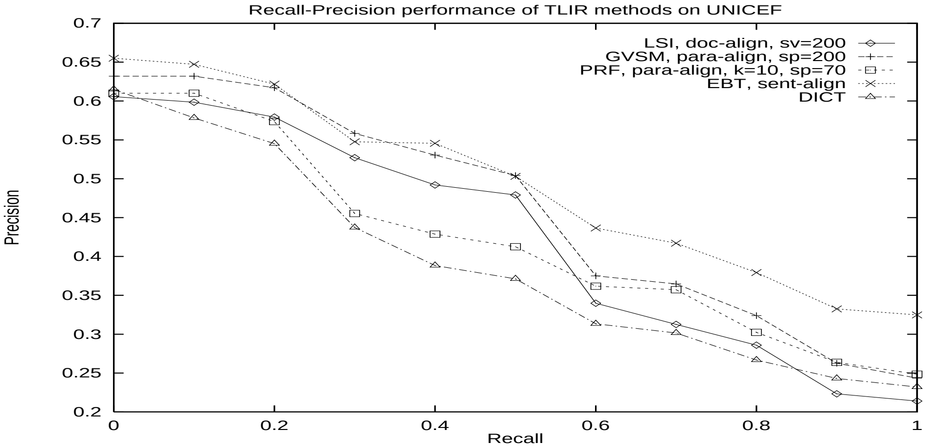

Figure 3: Recall-Precision performance of TLIR methods on UNICEF

For MLIR in the UNICEF corpus, we found that basic VSM, GVSM and LSI perform comparably best with an 11-point average precision (AVGP) of 0.49 to 0.50, whereas PRF performed worse with AVGP of 0.43 (This is not necessarily true for other corpora – see the results below on MEDLARS, where PRF outperforms VSM). We use the MLIR baseline (column 2 of table 1) to compare with TLIR performance in column 3 and to compute the AVGP ratio TLIR/MLIR as a percentage in column 4.

For TLIR, the performance of bilingual-dictionary term translation (DICT) was worst but still respectable at AVGP=0.39, corresponding to 80% of ML-VSM performance, and dictionary translation augmented with a large number of glossary entries (GLOSS) from the Pangloss MT project improved on DICT slightly yielding AVGP=0.41 (83%). EBT, in contrast, performed much better at AVGP=0.49, *slightly better* than ML-VSM. The two major reasons for the improvement of EBT over DICT and GLOSS are term frequency information and context-specific term translation (including an inherent query expansion described further in [2]), both derived automatically from the bilingual corpus. The query-expansion nature of the EBT should account for the surprise improvement over monolingual VSM, although this hypothesis requires testing by implementing an equivalent "back-translation"-based VSM query expansion and see if it produces a comparable improvement in ML-VSM.

All of the remaining translingual methods surpassed DICT, but none matched EBT's performance either in terms of absolute AVGP or in their ratio from monolingual performance, although they came close. GVSM exhibited better result in absolute performance (AVGP=0.46) than LSI (AVGP=.42) and PRF (AVGP=.42); PRF exhibited little degradation with a TL/ML ratio of 99%.

Different source-target text alignments were tested on the validation set for each corpus-based TLIR method except EBT (which always used sentence alignment), and sentence alignment proved best for PRF, while paragraph alignment proved best for GVSM and document alignment was optimal for LSI. Although TL-PRF performs at 99% of ML-PRF, its performance is quite sensitive to the value of K (the number of top-ranking documents for query expansion), as shown in Figure 4. If the user were willing to provide true relevance judgements, full relevance feedback should exhibit higher absolute performance, both for monolingual and translingual retrieval.

The early experiments reported in [7, 29] used the entire test set of 1121 documents, rather than dividing into validation and blind-test subsets reported in Table 1 and described in the previous paragraph. The results of the earlier experiments were similar, but slightly lower overall. In part, the improvements reported here are due to improved versions of all the methods (except PRF, which did not change).

For comparison, we also include translingual results reported by other researchers in Table 2. Because the methods have been run on different corpora with different queries, direct comparisons on absolute AVGP are not meaningful. However, the ratio of TLIR over MLIR results may be more indicative of the relative power of each TLIR method. The TLIR/MLIR degradation factors reported in the literature (primarily dictionary-based approaches) are comparable, though somewhat lower than our DICT and GLOSS methods: 40% to 73.5% versus our results of 80% to 83%. More interestingly, no previous results come close to the 87% to 101% TLIR/MLIR range exhibited by our corpus-based methods. We encourage direct comparisons on the same corpora in the future.

5.4 Effects of corpus alignment and parameter tuning in TLIR methods

We investigated different parameter values and different granularity alignments between the source and target language corpora, specifically at the document, paragraph and sentence levels.

For all of our TLIR methods, we used the *ntc* term weighting scheme ($TF * IDF$ with vector normalization) which appeared to be optimal on the UNICEF corpus.

For PRF, we first measured the effect of varying K (number of top-ranked documents used in query expansion) with different alignments. As shown in Figure 4, optimal performance for TL-PRF is at K=5 for document alignment and at K=10 for paragraph and sentence alignment, when tested on the validation set. (There are about 6 paragraphs per document and about 5 sentences per paragraph.) Therefore, we selected K=10 and paragraph alignment for our reported results on the test set. PRF performance is rather sensitive to K at the document-level alignment, but less sensitive (more robust) at the finer-grain alignments, hence the latter are preferable for stability as well as absolute performance reasons.

With K fixed at its optimal value for PRF, we tested how performance varies with changes in SP values (i.e., the number of TF*IDF top-ranking terms retained after query expansion). As shown in Figure 5, performance is rather stable and insensitive to SP larger than 30 on the validation set. Therefore, selected 70 as the SP value for our test. This modest expanded query size permits fast on-line performance.

GVSM has only one tunable parameter, i.e., the same SP as in PRF, but applied to both query and document vectors after the GVSM transformation. As shown in Figure 6, all alignments achieve stable performance in the validation set near the optimum with sufficiently large SP values, but document and paragraph alignments approach the performance plateau at much smaller SP values, resulting in faster on-line response and smaller storage for document indexes. We selected SP=200 for our blind test,

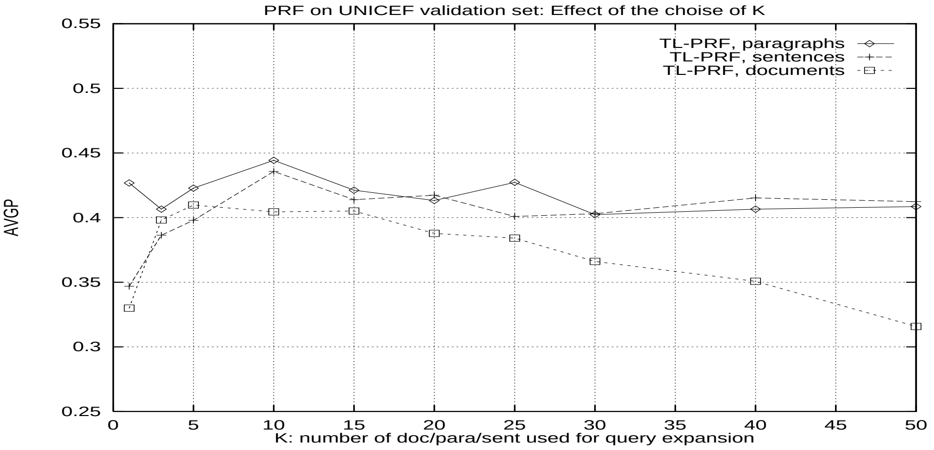

Figure 4: PRF performance w.r.t. parameter tuning on K

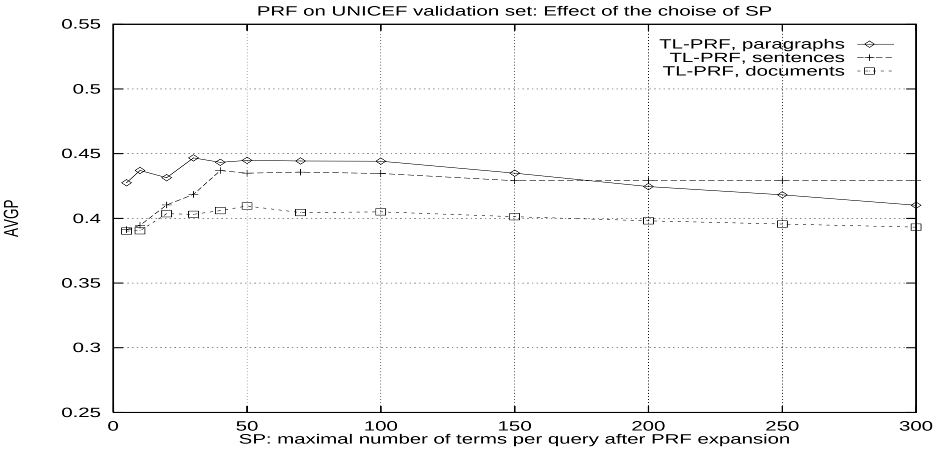

Figure 5: PRF performance w.r.t. parameter tuning on SP

although any value above 100 should perform comparably, according to our study on the validation set.

LSI also has a single tunable parameter, the number of singular values (SV) being used, which corresponds to the orthogonal dimensions of the reduced vector space, and is equivalent to the number of indexing terms per document or query after their LSI transformation. The performance of this method with different corpus alignment strategies is illustrated in Figure 7. Sentence-level alignment for LSI produces terrible results, both in terms of accuracy and computational time, and therefore is discarded from further consideration. Paragraph-level alignment also produces significantly worse retrieval results than document-level alignment does. The performance curve does not reach a plateau until 200 or more SVs are used for document alignment; the performance climbs slightly at SV values over 600. However, the computational cost increases superlinearly as with increasing SVs, and using SV=1000 defeats the original purpose of the SVD step in LSI: dimensionality reduction. Therefore, we selected SV=200, and document-level alignment for our blind test.

EBT has two tunable parameters: the filtering threshold used in generating the term dictionary and the total term weighting used in translating/expanding the query. For our experiments with the

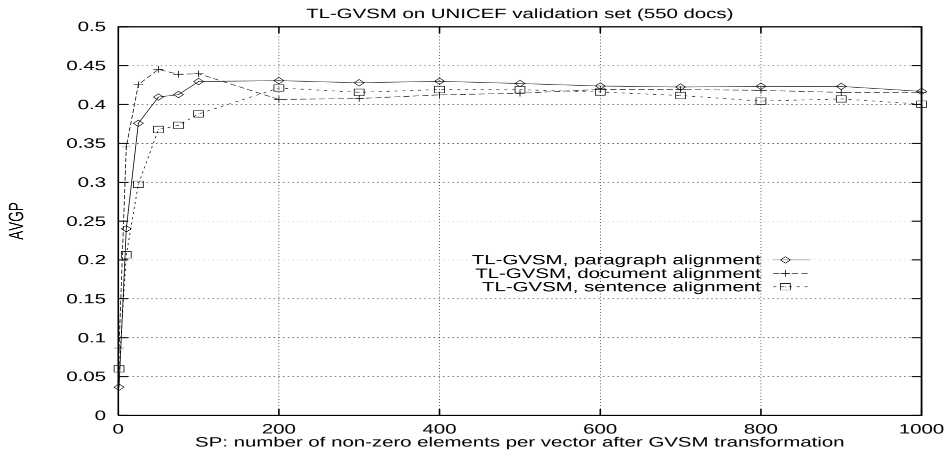

Figure 6: GVSM performance w.r.t. parameter tuning on SP

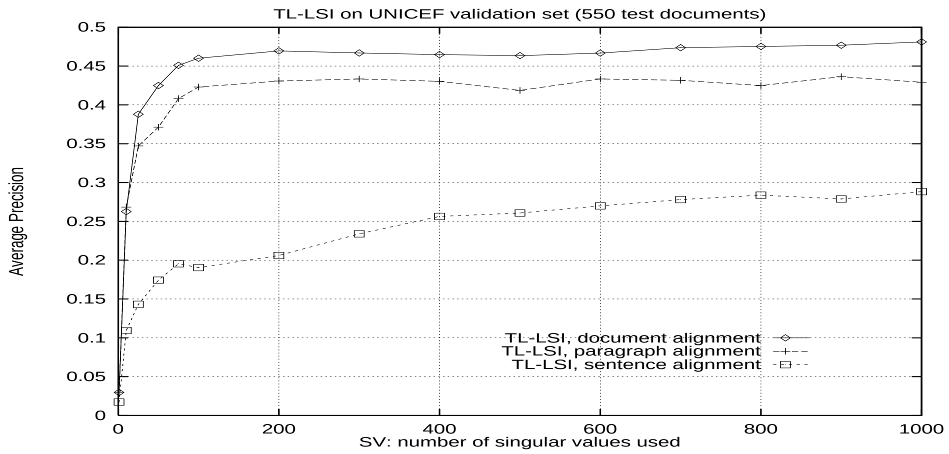

Figure 7: LSI performance w.r.t. to parameter tuning on SV

UNICEF corpus, EBT using the full UN multilingual corpus performed best with a filtering threshold of 0.27, i.e. the term dictionary consists of those word pairs which co-occur with each other in a translated sentence pair at least 27% of the time. A threshold of 0.11 performed best for a smaller corpus of twelve megabytes containing only the UNICEF training documents and four megabytes of non-UN texts which are also used in the full 250-megabyte corpus. For the blind test, therefore, the 12-megabyte corpus was used with a threshold of 0.11.

EBT term weighting is achieved by outputting multiples of each possible translation in proportion to the number of co-occurrences found when the term dictionary is generated, and allowing SMART to weight each word on the number of times it occurs in the newly-generated translated query. In our experiments, we found that performance does not change beyond a weighting of 20, i.e. a total of 20 words is output for each word in the query. For example, given the entry

```
(WATER (AGUA 1953)(ABASTECIMIENTO 753)),
```

EBT generates 14 occurrences of "agua" ($\frac{20\times1953}{1953+753}$, rounded) and 6 of "abasticimiento" for each occur-

Task	Method	ML.doc	TL.doc	ML.para	TL.para	Parameters
Mate	GVSM	.9897	.9352	.9897	.9573	SP=200,ntc
Mate	LSI	.9897	.9514	.9897	.9737	SV=200,ntc
Retr	GVSM	.3846	.3672	.4096	.4053	SP=200,ntc
Retr	LSI	.4286	.4148	.3916	.3834	SV=200,ntc

Table 3: Results summary of mate finding and retrieval

rence of "water" in the query. At least one occurrence of each word is output, even if the proportion rounds to zero.

5.5 Mate finding

LSI was first extended from MLIR to TLIR at Bellcore [17, 12], including the "fold-in" process mentioned earlier. However, the evaluation was unorthodox due to their lack of a bilingual corpus with queries and relevance judgements (such as the UNICEF corpus we prepared). *Mate-finding* was proposed: use a document in the source language as the query and determine if its translation (the "mate") is retrieved. Using LSI, a very high performance was achieved in 1990 and even higher in 1996. However, Dumais also reported that using machine translation produced an even slightly higher performance. The performance figures for mate-finding totally eclipse all published query-based document retrieval evaluations. We submit that the mate-finding task is far easier than true query-based retrieval, and thus good performance in the former may not be a meaningful indicator of performance in the latter. A document and its translation mate are extremely close - identical modulo translation in fact – unlike a query and the documents relevant to it. In order to test this hypothesis, we replicated the Bellcore mate-finding experiments on the UNICEF corpus, using both LSI and GVSM, and contrast those results with true retrieval in Table 3. We used the full set of 1121 test documents, i.e., the union of the validation test set (550 documents) and the evaluation test set (571 documents). The parameters are SV=200 for LSI, and SP=200 for GVSM, which are the optimal parameters found on the validation test set, as described in the previous section.

As expected, GVSM and LSI exhibit very high MLIR performance (both AVGP=0.99) and TLIR performance (AVGP=0.96 and 0.97 for paragraph alignment) in mate-finding. These are comparable to the results reported by Dumais *et al* in a different parallel corpus (the Canadian Hansard). But our corresponding query-based retrieval performance (AVGP=0.41 and 0.38 for paragraph alignment) differed extremely from mate finding. Hence, mate-finding does not reflect true IR performance in realistic tasks, and it should be discarded as an overly optimistic evaluation criterion for TLIR.

5.6 Monolingual retrieval on MEDLARS

We have shown the effectiveness of multiple corpus-based TLIR methods. However a question remains as to how intrinsically "difficult" the UNICEF corpus and queries are, compared to other (monolingual) corpora used in the IR community. To address this question we compared our monolingual IR results on UNICEF with the standard MEDLARS corpus and queries provided with the SMART system. Testing all our methods on each corpora also enabled us to see whether the relative performance ranking among the methods is preserved or not, in the monolingual case. Figure 8 presents the recall-precision curves for MEDLARS, showing:

- MEDLARS is an "easier" collection for IR, as shown by the fact that all the methods perform much better than in the UNICEF collection.
- All the corpus-based methods clearly outperform basic VSM, in MEDLARS, indicating that the use of empirical word associations and occurrence patterns provides significant benefits for MEDLARS. This is possibly the case because there is more room for improvement in terms of query expansion

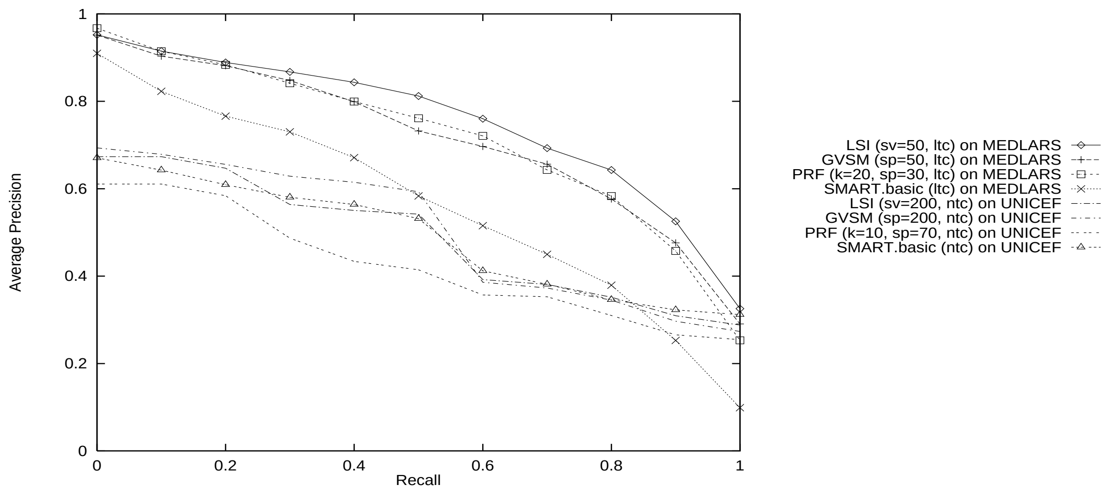

Figure 8: MLIR Performance of different methods on two corpora

for MEDLARS.

- Other than the basic VSM performance difference noted above, the relative ranking of the various methods did not change significantly. In both cases LSI performs better by a small margin over GVSM. PRF was somewhat closer to LSI performance in MEDLARS than in UNICEF.

The effectiveness of the corpus-based MLIR methods also depends on how clustered the relevant documents are (e.g., in PRF: whether the retrieved subset of the relevant documents is representative of the un-retrieved relevant documents), and how close the query is to the relevant documents. Therefore, for some corpora (such as MEDLARS), word expansion techniques (which are essential to our corpus-based methods) are more effective than for other corpora (such as UNICEF). Nevertheless, our central focus here is to cross the language barrier via learning from a bilingual corpus; we have found that all these methods are highly effective in solving this problem, regardless of how much they improve on (or degrade from) the baseline MLIR performance on a particular set of queries and documents.

6 Conclusions

This paper reports a thorough evaluation of multiple methods for translingual retrieval in a query-based retrieval task. Some methods were adapted from the literature and others are newly developed for TLIR. The latter set includes:

- *Example-Based Term Translation* – using a bi-lingual corpus to translate query terms in a corpus-relevant context.
- *Translingual Pseudo-Relevance Feedback* – using retrieved documents and their translations in a bi-lingual corpus for query formulation in the target language.
- *Translingual Generalized Vector Space Model* – using patterns of term occurrences in translated document pairs to establish translingual query-document similarities.

Our comparative study indicates that corpus-based methods clearly surpass methods based on general-purpose dictionaries, though results are a bit closer when the dictionaries are augmented with glossaries developed for Machine Translation systems. Our results demonstrate that TLIR methods can achieve performance approaching MLIR accuracy. More specifically, we conclude:

- Translingual retrieval is viable by a number of different techniques, ranging from term-based query translation and Pseudo-Relevance Feedback to Generalized Vector Spaces Model and latent Semantic Indexing.
- In our translingual retrieval test, Example-Based Term Translation performed best in absolute terms, but GVSM was a close second and LSI and PRF were not far behind. With respect to relative performance, all these methods showed only minor degradation from monolingual to translingual retrieval (TLIR/MLIR ratios of 87% to 101%).
- Dictionary-based query translation, though popular in the literature, should be re-examined as the TLIR method of choice given the results in this paper, though even there, our dictionary results, especially when enhanced with a glossary, performed acceptably.
- GVSM exhibited the most stable performance with respect to a large range of parameter values, while PRF had the smallest query-length after expansion for effective translingual retrieval, implying the fastest on-line response. LSI performance can reach a similar stability, but has a larger computation cost (time and space) in both the training and testing phases.
- Mate-finding is not a realistic test of translingual retrieval performance, when compared to standard evaluations with actual queries.

TLIR is quickly becoming a vibrant field, and this paper raises at least as many questions as it answers. Some follow directly from our work, and others follow from its limitations. Significant questions for future research include: Are there unexplored TLIR methods of comparable or better performance? Which methods scale to much larger collections, and at what cost in time and space? Which methods extend well to more disparate language pairs (such as English-Chinese)? Is it possible to exploit a small parallel corpus together with large monolingual ones? Can these methods be extended to *comparable corpora* (documents about the same topic in different languages, rather than translation mates)? Can machine translation play a more central role in TLIR, such as by automatically producing a parallel corpus for (part of) a collection? What should the user interface to a TLIR system be? Should it include MT of retrieved target-language documents?

Acknowledgments

We thank Christie Watson and Dorcas Wallace for their efforts in corpus annotation. We are also grateful to Xin Liu for his contributions to the improved and more efficient implementation of the GVSM and LSI methods.

References

[1] Lisa Ballesteros and Bruce Croft. Phrasal translation and query expansion techniques for cross-language information retrieval. In *20th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)*, pages 85–91, 1997.

[2] Ralf D. Brown. Automatically-Extracted Thesauri for Cross-Language IR: When Better is Worse. In *Proceedings of the First Workshop on Computational Terminology (COMPUTERM'98)*, August 1998.

[3] R.D. Brown. Example-Based Machine Translation in the Pangloss System. In *Proceedings of the Sixteenth International Conference on Computation Linguistics*, pages 169–174, 1996.

[4] R.D. Brown. Automated Dictionary Extraction for "Knowledge-Free" Example-Based Translation. In *Proceedings of the Seventh International Conference on Theoretical and Methodological Issues in Machine Translation*, 1997.

[5] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using smart: Trec 3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 69–80, 1995.

[6] J. G. Carbonell. New Approaches to Machine Translation. In *Proceedings of the conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Hamilton, NY, 1985.

[7] J.G. Carbonell, Y. Yang, R.E. Frederking, R. Brown, Y. Geng, and D. Lee. Translingual information retrieval: A comparative evaluation. In *Proceedings of IJCAI-97*, Nagoya, Japan, 1997. (Distinguished paper award).

[8] M. Davis and T. Dunning. A trec evaluation of query translation methods for multi-lingual text retrieval. In *The 4th Text Retrieval Conference (TREC-4)*, 1996.

[9] Mark Davis and William Ogden. Quilt: Implementing a large-scale cross-language text retrieval system. In *20th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)*, pages 92–98, 1997.

[10] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. In *J Amer Soc Inf Sci 1, 6*, pages 391–407, 1990.

[11] Ed. DK Harman. *Overview of the Third Text REtrieval Conference (TREC-3)*. US Government Printing Office, Washington, DC, 1995.

[12] S.T. Dumais, T.K. Landauer, and M.L. Littman. Automatic cross-linguistic information retrieval using latent semantic indexing. In *SIGIR'96 Workshop On Cross-Linguistic Information Retrieval*, 1996.

[13] R. Frederking, S. Nirenburg, D. Farwell, S. Helmreich, E. Hovy, K. Knight, S. Beale, C. Domashnev, D. Attardo, D. Grannes, and R. Brown. Integrating Translations from Multiple Sources within the PANGLOSS Mark III Machine Translation System. In *Proceedings of the first conference of the Association for Machine Translation in the Americas*, Columbia, MD, 1994.

[14] David Graff and Rebecca Finch. Multilingual Text Resources at the Linguistic Data Consortium. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*. Morgan Kaufmann, 1994.

[15] W. Hersh, C. Buckley, T.J. Leone, and D. Hickman. Ohsumed: an interactive retrieval evaluation and new large text collection for research. In *17th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*, pages 192–201, 1994.

[16] D.A. Hull and G. Grefenstette. Querying across languages: a dictionary-based approach to multilingual information retrieval. In *19th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96)*, pages 49–57, 1996.

[17] T. Landauer and M. Littman. Fully Automatic Cross-Linguage Document Retrieval using Latent Semantic Indexing. In *Proceedings of the 6th OED Conference on Text Research*, pages 31–38, 1990.

[18] M. Nagao. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. In A. Elithorn and R. Banerji (eds), editors, *Artificial and Human Intelligence*. NATO Publications, 1984.

[19] S. Nirenburg, J. G. Carbonell, M. Tomita, and K. Goodman. *Knowledge-Based Machine Translation*. Morgan Kaufmann Inc, San Mateo, CA, 1991.

[20] G. Salton. Automatic Processing of Foreign Language Documents. *Journal of American Society for Information Sciences*, 21:187–194, 1970.

[21] G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, Pennsylvania, 1989.

[22] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of American Society for Information Sciences*, 41:288–297, 1990.

[23] P. Sheridan and J.P. Ballerini. Experiments in multilingual information retrieval using the spider system. In *19th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96)*, pages 58–65, 1996.

[24] P. Sheridan, M. Wechsler, and P. Schauble. Cross-language speech retrieval: establishing a baseline performance. In *20th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)*, pages 99–107, 1997.

[25] Padmini Srinivasan. Optimal document-indexing vobabulary for MEDLINE. *Information Processing & Management*, 32(5):503–514, 1996.

[26] S.K.M. Wong, W. Ziarko, V.V. Raghavan, and P.C.N. Wong. On modeling of information retrieval concepts in vector space. In *ACM Transaction of Database Systems*, number 2, pages 299–321, 1987.

[27] S.K.M. Wong, W. Ziarko, and P.C.N. Wong. Generalized vector space model in information retrieval. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'85)*, pages 18–25, 1985.

[28] Y. Yang. Noise reduction in a statistical approach to text categorization. In *Proceedings of the 18th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95)*, pages 256–263, 1995.

[29] Y. Yang, R.E. Frederking, J.G. Carbonell, R. Brown, Y. Geng, and D. Lee. Bilingual-corpus based approaches to translingual information retrieval. In *Proceedings of MULSAIC-97*, Nagoya, Japan, 1997.

[30] Y. Yang and J.P. Pedersen. Feature selection in statistical learning of text categorization. In *The Fourteenth International Conference on Machine Learning*, pages 412–420, 1997.