

## **Multi-lingual Information Retrieval**

*Ted Dunning*

*Mark Davis*

Computing Research Laboratory  
New Mexico State University  
Las Cruces, NM

### *ABSTRACT*

We have designed a fully multi-lingual information retrieval system and tested crucial parts. This system can accept a query in one language and find documents in others. Furthermore, relevance feedback can be used in a fully multi-lingual fashion.

Our system is based on the availability of parallel and aligned texts. We use these texts to derive a linear approximation of the translation process, and then use this linear transformation to implement a conventional vector based information retrieval system. We describe three possible techniques for deriving this translation matrix, one of which we have implemented and tested on a relatively moderately sized training corpus. Our method appears to be very efficient in terms of the size of the necessary training corpus.

Since our solution for the translation matrix is incremental in nature, additional parallel texts can be used to augment the system at any time.

February 5, 1993

# Multi-lingual Information Retrieval

*Ted Dunning*

*Mark Davis*

Computing Research Laboratory  
New Mexico State University  
Las Cruces, NM

## Introduction

Information retrieval is one of the most successful areas of natural language processing. This success has been largely due to the fact that the words which appear in a document generally suffice to determine the subject matter. This means that deep semantic processing is not necessary to produce relatively good results and that very large collections of texts can be searched very quickly. Since the late 60's, mono-lingual text retrieval systems have been available with all the essential features of the most effective systems available today.

Text retrieval in a multi-lingual setting also has a long history, although critical technology has been lacking until recently. Salton discussed the possibility of multi-lingual text retrieval as long ago as 1970, but apparently did not actually construct a system [Salton, 1970]. In this proposed system, it was expected that a bi-lingual thesaurus would be used to translate query terms to be used by a traditional query system. Unfortunately, systems based on thesauri are limited by the sorts of thesauri which are available. In particular, domain specific dictionaries and thesauri are difficult to come by and are largely obsolete by the time they are available. Furthermore, dictionaries capture a particular form of lexical knowledge which is often very different from the sort needed to specifically relate words or phrases.

Experiments with latent semantic indexing [Landauer and Littman, 1990] showed that paragraphs which were translations of each other could be retrieved but again no actual retrieval system was built, nor was it clear how the system would perform in practice. This use of parallel corpora eliminates many of the problems of using bi-lingual dictionaries, but introduces new problems. In particular, in the context of a traditional vector based retrieval system, it has not been clear how to perform multi-lingual retrieval based on the information contained in parallel translated corpora. The success of experiments done with Latent Semantic Indexing does not directly provide a method to make a more traditional vector based system work.

We have developed a system for doing multi-lingual text retrieval which is based on a novel method for solving very large systems of linear equations. This system views vector based text retrieval as an exercise in computing inner products, and views translation as a linear transformation of a feature vector. This transformation can be applied to either the feature vectors derived from documents, or it can be applied to the

feature vector of the query being used to retrieve the documents. This duality means that the transformation can be derived by relating feature vectors taken from a parallel corpus, but applied to queries such as are used in traditional retrieval systems.

### **Previous use of Parallel Corpora**

Much of the current interest in the use of parallel corpora stems from the ongoing research being conducted by the IBM speech recognition group at IBM's T.J. Watson laboratories. This group first popularized the use of the Hansard records of Canadian parliamentary debate. They also developed many of the statistically based methods for the analysis of text. Their work on statistical part of speech tagging has been replicated by many people including Church [Church, 1988], researchers at BBN [Meteer, Schwartz and Weischedel 1991] and others. Their work on the alignment of sentences based on sentence lengths was also replicated with some changes by Gale and Church [Gale and Church, 1991]. More recently they have had good luck on the resolution of sense ambiguities as evidenced by alternative translations [Brown, et al, 1991]. This work has also been replicated by Gale and Church [Gale and Church, 1992].

All of these developments by the IBM group are part of an effort to apply the basic signal estimation paradigm that they have used for speech recognition to the task of translating French into English. As such, they have expended considerable effort on the problem of aligning words in parallel sentences and on the problem of modeling English in isolation.

We have found that the needs of an information retrieval system are considerably simpler than the needs of a translation system and that this allowed substantial simplification of the mathematical systems needed to derive information from parallel corpora. In traditional document-level retrieval systems, both the query and document are considered unordered "bags of words". A translation system which produces unordered sets of translated words for a given document should therefore provide adequate cross-language retrieval. Requiring only unordered translations means that all considerations of order and word alignment could be eliminated from our system, substantially simplifying the translation process. Much of the complexity of the methods used by the IBM group [Brown, et al. 1993] are directly due to the added complexity of considering word order.

### **Vector based text retrieval**

Treating both documents and queries as vectors containing weighted word frequencies as the vector elements is the basis for substantial work in information retrieval [Salton 1975, 1979, 1983]. In this approach, the inner vector product of the feature vectors is used as a measure of similarity between the document and query vectors. Best results are typically obtained by normalizing the document and query vectors, giving what is normally called a cosine measure. The value in the  $i$ th position in the document and query vectors are weighted counts of the number of times word  $i$  occurs in the document or query respectively.

### **Linear transformation based translation**

It is possible to view the process of translation as an approximation of a linear transformation if we only look at the translation of relatively long strings. This is true because the translation of two strings is usually the concatenation of their translations. While this linearity breaks down dramatically at the word level, at the sentence level and above, it works fairly well.

In the absence of the normalization required by the cosine distance measure, conversion to feature vectors is also a linear process in that the concatenation of texts results in the addition of the corresponding feature vectors.

Pushing this linear metaphor further, we can solve for the linear transformation of document vectors which gives us the minimum squared error. The resulting linear transformation will allow the free translation of feature vectors of documents and queries, but because the conversion of a document into a feature vector is not invertible, this linear transformation cannot be used for translation.

Our approach is in strong contrast to the approach taken by the group working on statistically based machine translation at IBM. They prefer to use a metaphor taken from communication theory in which the observed string (say, in French) is taken to be the result of the translation of an original string (say, in English). Their problem then is the derivation of this original string. This inversion of the normal model of translation allows the use of several very powerful tools based on signal estimation methods. Unfortunately, the search for the string which has maximum likelihood of being the source of the observed string is computationally quite expensive. The method we have developed is able to 'translate' a query at the cost of one (large) matrix multiplication.

The work reported by Landauer and Littman [Landauer and Littman, 1990] provides another interesting illustration of the potential linearity of the translation process. In this work, a matrix is formed whose rows represent the word counts in each document and thus are quite comparable to our feature vectors. They derive an approximate basis for the vector space spanned by the rows of this matrix by computing the singular value decomposition of this matrix. In the bilingual case, they adjoin the matrix formed by the English version of the Hansard corpus with the matrix formed by the French version of the corpus. They then compute an approximate basis for the adjoined matrix and use this basis for retrieval of paragraphs in either language. Deriving this single basis for the adjoined word count matrix not only would allow retrieval as described in the Landauer and Littman paper, but in fact should allow a translation matrix to be derived directly from the adjoined basis. We have yet to explore this possibility.

The difficulty with such an approach lies in the computational resources required to compute the necessary singular value decomposition. As part of their participation in the recent Text Retrieval Evaluation Conference, the singular value decomposition of a small subset of the test texts was computed. This effort was conducted on a workstation equipped with nearly 400MB of main memory. Interestingly, our programs ran on machines with approximately one tenth this amount of memory and were considerably

simpler (albeit, our system attacked somewhat smaller problems). It is an interesting and open question whether or not the matrix computed by our program could be used to derive an approximate basis for the word count matrix.

### Multi-lingual text retrieval

Given a parallel corpus of natural language texts consisting of a large number of short texts in two languages which happen to be paired so that the elements of the pairs are translations of each other, it is possible to derive enough information to build a bilingual information retrieval system. To this end, it is not necessary to actually build a translation system along the lines of IBM's statistically based system. Instead, we can build a system which translates the feature vectors which are used by the basic retrieval engine. By using several such parallel corpora, we can build a truly multi-lingual system.

For concreteness in the following description, assume that we have texts in English and Japanese. We will write the feature vectors of these texts as  $E_i$  and  $J_i$  respectively where  $i$  ranges over the various documents. In different systems, these features represent different things. In many systems, they are derived from word counts, in others they may represent something very different, such as coordinates in the synthetic feature space derived by using latent semantic indexing.

Whatever the features represent, for a variety of vector scoring systems including those based on cosine scores, latent semantic indexing, as well as many probabilistic systems, the score can be written as an inner product of a query feature vector and a document feature vector. Either or both of the query or the feature vector may be normalized ahead of time as in the cosine scoring system used in the SMART retrieval system, but the basic mechanics of the inner product still apply. Thus, the score which is assumed to reflect the relevance of a particular document  $E_i$  to the query  $Q$  is simply  $Q^T E_i$

If we could associate with any English query  $Q$ , a Japanese query  $Q'$  such that the equivalence

$$Q^T E_i = Q'^T J_i$$

held, perhaps by relating  $Q$  and  $Q'$  via a linear transformation  $T$ , then we could construct a bilingual text retrieval system because retrieval of Japanese documents could be done by computing  $Q'$  and using that to retrieve Japanese documents. Introducing this linear transformation gives the equations

$$Q^T E_i = (T Q)^T J_i$$

This can easily be rearranged so that the documents are translated instead of the query,

$$Q^T E_i = Q^T (T J_i)$$

This makes it clear that we don't really need to introduce the query at all to derive  $T$ . In particular, any solution of the system

$$E_i = T J_i$$

will also be a solution of the system which includes the query. It is important to note that  $T$  can be derived directly from the documents, but at query time can be applied to the query. In practice, the system above will be inconsistent, but by successively adjusting  $T$  so that each of the equations in the system is true, it is possible to converge to a usable approximation.

### Solving for the linear transformation

An iterative method for solving for  $T$  can be derived by solving

$$E_i = (T + \Delta T) J_i$$

for the  $\Delta T$  that minimizes  $\sum_{ij} \Delta T_{ij}^2$

The solution to this constrained minimization problem is

$$\Delta T = \frac{(E_i - T J_i) J_i^T}{J_i^T J_i}$$

There are a wide variety of software packages which were designed to solve sparse linear systems. Unfortunately, these packages are not directly suitable for this problem since they are generally designed to solve systems of the form

$$A x = b$$

where the unknown vector  $x$  and the constant vector  $b$  are dense vectors while  $A$  is a sparse matrix. In addition, these packages are designed to handle problems where  $A$  is full rank and the length of  $x$  and  $b$  is no more than a few tens of thousands. Our system, on the other hand, is of the wrong form, and is inconsistent. Even in manufactured examples which are consistent, the system is typically massively underdetermined.

We have successfully implemented this iterative method of solution and it appears to converge relatively quickly to a usable solution. In fact, initial convergence appears to occur in less than one pass through the training set of documents. This is apparently because common words appear often enough and in varied enough settings that the system is able to determine those portions of  $T$  which describe the translation of those common words. As  $T$  comes to describe these common words, the error term  $(E_i - T J_i)$  more and more will be significantly non-zero when unusual words appear. Since these unusual words will (almost by definition) appear in isolation,  $\Delta T$ , the update to  $T$ , will handle these new words in one step. From the standpoint of numerical analysis, this convergence in less than a single epoch is highly unusual.

Another as yet unimplemented option for solving for  $T$  is to directly attempt to minimize

$$\mathbf{R} = \sum_{ij} (E_{ij} - \sum_k T_{ik} J_{kj})^2$$

By taking the derivatives with respect to all the various  $T_{ab}$ , we get the equations

$$\sum_k (\sum_j J_{bj} J_{kj}) T_{ak} = \sum_j E_{aj} J_{bj}$$

At first glance, this appears to be an enormous number of equations. But, in fact, these equations can be solved for one row of  $T$  at a time, considerably easing the computational burden. It remains an open question whether or not the coefficient matrix for each row is sparse or not. This method has the strong advantage of solving directly for the desired minimum, rather than iteratively approaching the minimum, and if the coefficient matrix is sparse then available sparse matrix solution packages such as SPARSE can be used. Since SPARSE is actually part of the widely used Spice circuit simulation package, considerable efforts have been expended to make SPARSE run as efficiently as possible. We plan to investigate this method further.

### **Practical Implementation of the Update Method.**

In our implementation of the iterative solution method, we define a matrix as an array of rows, each of which is a sparse vector. Each of these sparse vectors is an array of structures containing an index and a value. The elements of structures are sorted by index to simplify operations. Typically, when a vector is created or expanded, new elements are added to the end of the vector without regard to index order, and once a number of additions have been made, the vector is resorted. When all available storage for a vector is exhausted, new storage twice the size of the previous vector is allocated and all elements of the vector are copied to the new area. The amortized cost of the copy, sort and reallocation is very small.

This method of storing sparse vectors avoids the storage and complexity overhead of methods which would keep the vectors in index order, and does not hurt performance significantly; less than 10% of program run time is spent sorting. Increasing the storage overhead of a vector, on the other hand, would likely seriously impact run time due to paging effects.

In order to avoid excessive fill-in of  $T$ , we only store the  $N$  largest elements at each iteration. This trimming is easily done by sorting the vector in descending order by magnitude, setting the number of active elements to  $N$ , and then resorting by index. The effect of this trimming is to substantially decrease storage requirements, at the cost of increasing the number of iterations which must be done to find convergence. On the other hand, this trimming also improves the stability of the convergence. Currently, we use  $N = 30$ , based on the assumption that most words have considerably fewer than 30 translations.

### **Operational Details**

We have run the iterative solution method using two parallel corpora. One is a collection of 10,000 abstracts in Japanese and English from the Japanese Institute of Computer Science Technology (JICST) and the other is a collection of about 100,000 words taken from a total of 1.31 million in Spanish and English of reports from the Pan American Health Organization (PAHO). The JICST corpus was unfortunately a mixture of translations, summaries and independent abstracts, so it is not practical to sentence align it. The PAHO corpus, on the other hand, consists solely of translations of documents. We used the portion of the PAHO corpus that had previously been aligned

for another CRL project. Alignment was done using a variation of Church's algorithm (Church and Gale 1990).

For the JICST corpus, the English text contained just under 1.4 million words total and 60,000 unique words. For the Japanese abstracts, contiguous strings of katakana, hiragana and kanji were each considered to be "words" for tokenization purposes, resulting in about 90,000 unique Japanese words. A reasonable alternative would be to consider each kanji character a word, although this possibility has not been explored.

The PAHO parallel corpus demonstrated rapid convergence, with total squared error dropping to approximately 30% of the initial error levels in 194 training epochs. An epoch is one complete cycle of iterative adjustment through the training set and requires about 20 minutes on a Sun/4 with 64 Mb of memory. Most of the convergence takes place in the first 20 epochs, although slow improvement continued as long as the program continued.

Convergence for the JICST corpus was substantially slower due to the greater size of the vocabulary and the noisy nature of the corpus as a whole. A single epoch took over 1 day to complete and convergence appeared slower.

We attribute the slower convergence and running time of the JICST corpus primarily to the total vocabulary, the larger number of words in each aligned segment of text (100-3000 vs. 1-40) and the considerably higher difficulty in developing a translation matrix from texts which are sometimes not even reasonable paraphrases of each other.

Overall, the program required memory space proportional to the number of unique words in the English text. The translation matrix was limited to no more than 30 non-zero elements per row, resulting in an active data set requirement of about 250 bytes per row. For the PAHO corpus, this means that the matrix occupies about 2.5MB of memory. For the JICST corpus, about 25-30MB were required. In addition, it is helpful to have enough memory to cache a substantial number of the training set vectors. For the JICST corpus, this means that about another 10-15MB of storage is needed. Since total vocabulary is not likely to increase much with corpus size, we expect this method to work well on much larger corpora.

## Empirical Results

We have performed several initial tests on the quality of the translation matrix. The most indicative of the quality of translation is shown in Figure 1. 10 novel paragraphs were chosen from the English half of the PAHO corpus and their feature vectors were computed using the same weighting scheme as in the training texts. In addition, the corresponding Spanish paragraphs were converted to feature vectors and translated using the translation matrix derived from the training texts. If we label the English vectors  $E[i]$  and the corresponding Spanish vectors  $S[i]$ , Figure 1 shows a scatter plot of  $E[i]^T E[j]$  versus  $T S[i]^T E[j]$ . As can be seen, there is a strong correlation between the results obtained using the English paragraphs and the translated Spanish paragraphs. The fact that the points fall for the most part below the ideal line indicates that novel

words were occurring which were not accounted for in the translation matrix  $T$ .

That the translation matrix performed this well on such a tiny training corpus bodes well for the success of this method for information retrieval in that it indicates that the relevancy scores obtained using a Spanish query and an English query should be very nearly the same.

Another interesting test is to print out the translation matrix in various forms. In the Table 1, 20 randomly selected rows and 20 randomly selected columns of the translation matrix are displayed. The rows are displayed by showing the English word associated with the row followed by the 5 Spanish words corresponding to the largest entries in that row (in descending order) and the columns are displayed by showing the Spanish word corresponding to the column followed by the 5 English words which represent the 5 largest entries in that column. As can be seen, the correct translation occurs first a significant portion of the time, and the correct translation occurs in the top five words a great majority of the time. Although not shown, similar tables can be constructed by examining  $T^{-1}$  (computed by reversing the roles of the two languages). Some of the translations shown in the table are rather idiosyncratic and particular to the training texts, but the fact that such generalizations can be made after only about 50,000 words in Spanish and English is somewhat surprising. Other statistical translation systems have involved training data sets three orders of magnitude larger.

The most important test of the method, however, is to incorporate it into a working IR system and test it using real data and users.

## Conclusions

Based on this preliminary work, it appears that parallel corpora can be used to construct a practical multi-lingual information retrieval system. There are serious unresolved questions about the relative merits of various solution methods for finding the required translation matrix, but it is clear that such matrices can be found within the limits of current computers. It is also clear that the current results do not give an adequate picture of how the use of a translation matrix will affect overall retrieval system performance.

We plan to continue work in an effort to answer these and other questions, as well as to build practical systems based on this technology. For information on the availability of the software and training data used in this work, contact the Consortium for Lexical Research via email at [lexical@nmsu.edu](mailto:lexical@nmsu.edu), or the authors at [ted@nmsu.edu](mailto:ted@nmsu.edu).

evaluation	evaluación	dio	problemática	complementaria	basada
into	incluye	Integrar	ii	brindó	salubridad
ends	epidémico	manifestar	latido	extingue	deja
services	servicios	proyecto	modelo	32	gobernadores
strengthening	fortalecimiento	de	garantizar	técnico	contenido
employees	empleados	esto	debido	Ocupacional	fuerza
well-being	orienta	incapacitantes	accidente	sólo	productividad
concept	concepto	su	HACCP	Concepto	microbiológica
alignment	refiere	masiva	favorecer	reconocer	formar
enable	Así	coordinada	activos	pre-inversio-	nes
priority	registros	bibliográficos	analizar	recopilar	1800
both	heces	campos	tales	propios	mayoría
cost	detectá	ninguna	odontológica	diferencia	otra
encouraging	publicar	apoyar	adecuada	desagregadas	Recomendaron
produce	desafíos	vectores	enfrentando	andinos	520
Center	Centro	Metepec	317	humanos	FORTALECIMIENTO
Education	Ciencias	junio	I	UDUAL	Habana
basically	Mundial	maternoinfantil	Favor	Niño	Nacional
various	es	diversos	servicios	destinado	rama
diagnóstico	diagnosis	status	treatment	cooperate	providing
generales	general	below	considerations	PTCs	eighth
emigraron	result	refugees	reentry	persons	war
poblaciones	populations	collaborated	executing	increasing	solve
bienestar	well-being	useful	families	opportunity	handicapped
Subcomité	Subcommittee	emphasis	Women	disseminating	special
muestra	sample	characterizes	heterogeneity	fresh-water	eleven
vacunación	remain	vaccination	adolescence	sweeping	inexpensive
actualizada	updated	PAHO	listing	preparing	Library
representativos	titles	represent	listings	unsolicited	acquired
anteriormente	addition	achieve	previously	incomes	subcomponents
febrero	wild	eradication	managerial	transmission	pedagogical
prestan	broad	Other	meet	oriented	improving
realizados	had	involved	changes	These	selected
búsqueda	284	authorities	collaborated	disease	prevention
manipulación	RISK	undesirable	euthanasia	handling	orienting
establezcan	obligations	transactions	allotments	write-offs	Recording
partidos	framework	clear	makes	consensus	consolidated
PBI	foreign	investment	satisfactory	favorable	GDP
río	All	River	located	south	Abajo

Table 1. Sample word translations from translation matrix