Automatic Cross-Language Information Retrieval using Latent Semantic Indexing

Michael L. Littman
Susan T. Dumais
Thomas K. Landauer

October 7, 1996

1 Introduction

We describe a method for fully automated cross-language document retrieval in which no query translation is required. Queries in one language can retrieve documents in other languages (as well as the original language). This is accomplished by a method that automatically constructs a multi-lingual semantic space using Latent Semantic Indexing (LSI). We present strong preliminary test results for our cross-language LSI (CL-LSI) method for a French-English collection. We also provide some evidence that this automatic method performs comparably to a retrieval method based on machine translation (MT-LSI).

For the CL-LSI method to be used, an initial sample of documents is translated by humans or, perhaps, by machine. From these translations, we produce a set of dual-language documents (i.e., documents consisting of parallel text from both languages) that are used to "train" the system. An LSI analysis of these training documents results in a dual-language semantic space in which terms from both languages are represented. Standard mono-lingual documents are then "folded in" to this space on the basis of their constituent terms. Queries in either language can retrieve documents in either language without the need to translate the query because all documents are represented as language-independent numerical vectors in the same LSI space.

We compare the CL-LSI method to a related method in which the initial training of the semantic space is performed using documents in one language only. To perform retrieval in this single-language semantic space, queries and documents in other languages are first translated to the language used in the semantic space using machine translation (MT) tools. We found that both this MT-LSI method and the CL-LSI method performed extremely well in two test retrieval scenarios.

2 Overview of Latent Semantic Indexing (LSI)

Latent Semantic Indexing is a variant of the vector-retrieval method [12] in which the dependencies between terms are explicitly modeled and exploited to improve retrieval. One

advantage of the LSI representation is that a query can retrieve a relevant document even if they have no words in common.

Most information-retrieval methods depend on exact matches between words in users' queries and words in documents. Typically, documents containing one or more query words are returned to the user. Such methods will, however, fail to retrieve relevant materials that do not share words with users' queries. One reason for this is that the standard retrieval models (e.g., Boolean, standard vector, probabilistic) treat words as if they are independent, although it is quite obvious that they are not. A central theme of LSI is that term-term inter-relationships can be automatically modeled and used to improve retrieval; this is critical in cross-language retrieval since direct term matching is of little use.

LSI examines the similarity of the "contexts" in which words appear, and creates a reduced-dimension feature-space representation in which words that occur in similar contexts are near each other. That is, the method first creates a representation that captures the similarity of usage (meaning) of terms and then uses this representation for retrieval. The derived feature space reflects these inter-relationships. LSI uses a method from linear algebra, singular value decomposition (SVD), to discover the important associative relationships. It is not necessary to use any external dictionaries, thesauri, or knowledge bases to determine these word associations because they are derived from a numerical analysis of existing texts. The learned associations are specific to the domain of interest, and are derived completely automatically.

The singular-value decomposition (SVD) technique is closely related to eigenvector decomposition and factor analysis [3]. For information retrieval and filtering applications we begin with a large term-document matrix, in much the same way as vector or Boolean methods do. This term-document matrix is decomposed into a set of k, typically 200–300, orthogonal factors from which the original matrix can be approximated by linear combination; this analysis reveals the "latent" structure in the matrix that is obscured by noise or by variability in word usage.

Figure 1 illustrates the effect of LSI on term representations using a geometric interpretation. Traditional vector methods represent documents as linear combinations of orthogonal terms, as shown in the left half of the figure, so that the angle between two documents depends on the frequency with which the same terms occur in both, without regard to any correlations among the terms. Here, Doc 3 contains Term 2, Doc 1 contains Term 1, and Doc 2 contains both terms. In contrast, LSI represents terms as continuous values on each of the k orthogonal indexing dimensions. Since the number of factors or dimensions is much smaller than the number of unique terms, terms will not be independent as depicted in the right half of Figure 1. When two terms are used in similar contexts (documents), they will have similar vectors in the reduced-dimension LSI representation. LSI partially overcomes some of the deficiencies of assuming independence of words, and provides a way of dealing with synonymy automatically without the need for a manually constructed thesaurus. (Earlier work [4, 6] presented detailed mathematical descriptions and examples of the underlying LSI/SVD method.)

The result of the SVD is a set of vectors representing the location of each term and document in the reduced k-dimension LSI representation. Retrieval proceeds by using the terms in a query to identify a point in the space—technically, the query is located at the

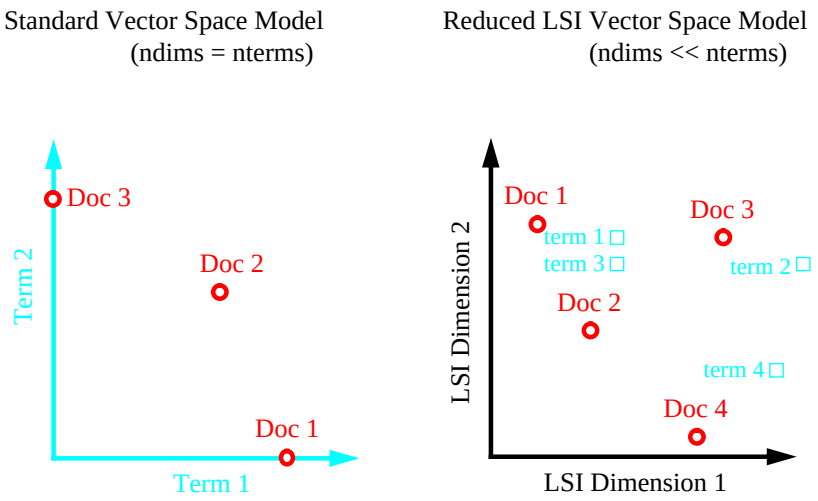

Figure 1: Term representations in the standard vector vs. reduced LSI vector models.

weighted vector sum of its constituent terms. Documents are then ranked by their similarity to the query, typically using a cosine measure of similarity. While the most common retrieval scenario involves returning documents in response to a user query, the LSI representation allows for much more flexible retrieval scenarios. Since both term and document vectors are represented in the same space, similarities between any combination of terms and documents can be easily obtained—one can, for example, ask to see a term's nearest documents, a term's nearest terms, a document's nearest terms, or a document's nearest documents. We have found all of these combinations to be useful at one time or another.

New documents (or terms) can be added to the LSI representation using a procedure we call "folding in." This method assumes that the LSI space is a reasonable characterization of the important underlying dimensions of similarity, and that new items can be described in terms of the existing dimensions. Any document not used in the construction of the semantic space is located at the weighted vector sum of its constituent terms. This is exactly how queries are handled and has the desirable mathematical property that a document that is already in the space is folded in at the same location (i.e., it receives the same representation). A new term is located at the vector sum of the documents in which it occurs.

In single-language document retrieval, the LSI method has equaled or outperformed standard vector methods in almost every case, and was as much as 30% better in some cases [4, 5].

3 Cross-Language Retrieval Using LSI

The LSI retrieval method has performed well in mono-lingual (English) document collections. We will now describe the CL-LSI method, in which LSI is adapted to cross-language retrieval.

Hon. Erik Nielsen (Deputy Prime Minister and Minister of National Defence): Mr. Speaker, we are in constant touch with our consular officials in Libya. We are advised the situation there is stabilizing now. There is no immediate threat to Canadians. Therefore my responses yesterday, which no doubt the Hon. Member has seen, have not altered.
L'hon. Erik Nielsen (vice-premier ministre et ministre de la Défense nationale): Monsieur le Président, nous sommes en communication constante avec nos représentants consulaire en Libye. D'après nos informations, la situation est en train de se stabiliser, et les Canadiens ne sont pas immédiatement menacés. Par conséquent, mes réponses d'hier, dont le représentant a dû prendre connaissance, n'ont pas changé.

Table 1: A dual-language document used in training the CL-LSI system.

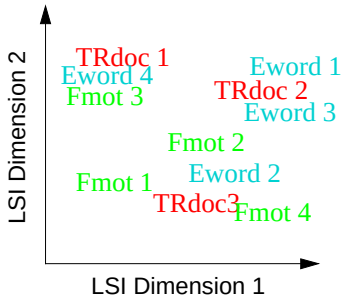

Figure 2: The training phase of CL-LSI. Each training document is a pair of English and French versions of the same document.

3.1 The Basic Idea

An initial sample of documents is translated by human or, perhaps, by machine, to create a set of dual-language training documents. An example of such a training document from the Hansard collection (the Canadian Parliament proceedings) used in our experiments is given in Table 1. The LSI method ignores word order and, therefore, treats this document as a bag of freely intermingled French and English words.

A set of training documents like this is analyzed using LSI, and the result is a reduced dimension semantic space in which related terms are near each other as shown in Figure 2. The space contains the training documents (TRdoc 1 through TRdoc 3 in the figure), and the terms that appear in the training documents. Because the training documents contained both French and English terms, the LSI space will contain terms from both languages (Eword 1 through Eword 4 in English and Fmot 1 through Fmot 4 in French); this is what makes it possible for the CL-LSI method to avoid query or document translation. Words that are consistently paired in translation (e.g., Libya and Libye) will be given identical representations in the LSI space, whereas words that are frequently associated with one another (e.g., not and pas) will be given similar representations.

The next step in the CL-LSI method is to add (or "fold in") documents in just French or

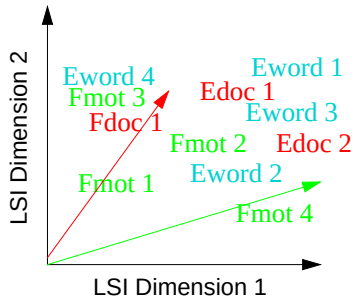

Figure 3: The fold-in and query phases of CL-LSI. Monolingual documents are located at the vector sum of their constituent terms.

English as depicted in Figure 3. This is done by locating a new document at the weighted vector sum of its constituent terms (e.g., Edoc 1 in the figure). The result of this process is that each document in the database, whether it is in French or in English, has a language-independent representation in terms of numerical vectors. Users can now pose queries in either French (black vector in Figure 3) or English (grey vector in Figure 3) and get back the most similar documents regardless of language.

3.2 Experimental Tests

Landauer and Littman [9] described some simple retrieval experiments using CL-LSI applied to the Hansard collection. They worked with a sample of 2,482 English paragraphs and the same 2,482 paragraphs in French. These paragraphs were selected by sampling the Hansard collection from 1986 to 1989 and retaining only paragraphs that contained at least five lines in both the English and French versions. The "documents" averaged 84 words in English and 86 words in French; thus, the combined training documents averaged 170 words.

Using the same document collection, we replicated some of the results from the earlier study and began a battery of new tests. We randomly divided the 2,482 documents into a training set of 982 dual-language documents and a test set consisting of 1,500 English documents and their 1,500 corresponding documents in French. As described in the previous section, the 982 dual-language documents were used to create a dual-language semantic space. (On a standard Sparc workstation, this type of analysis takes less than 2 minutes.) We used a 982-dimensional representation of this space—this is the largest possible given the training set and we found that this tended to improve the results over using fewer dimensions.

The 1,500 French-only test documents and 1,500 English-only test documents were then folded in to the dual-language space. As a result, each of these documents was assigned a 982-dimensional language-independent representation.

3.2.1 Finding a Cross-language Mate

Ideally, we could use a standard multi-language test collection with cross-language queries and relevance judgments to evaluate the CL-LSI retrieval system. As we are unaware of any generally available collection at this time, we were forced to create alternative tests where the relevance of documents to queries could be derived automatically.

query	document	returned mate at rank 1	percentage
English	French	1,475	98.3%
French	English	1,478	98.5%

Table 2: Results for cross-language mate finding using CL-LSI.

query	document	returned mate at rank 1	percentage
English	French	715	47.7%
French	English	743	49.5%

Table 3: Results for cross-language mate finding using the ordinary vector method (not LSI).

The first test uses documents to find their cross-language mates and can be thought of this way. Treat each of the 1,500 English documents as queries, each with exactly one relevant document in French—its translation (or mate). Now ask, for how many of the queries is the relevant document returned first? The results are presented in Table 2, which show that the CL-LSI method does an excellent job of assigning translations similar representations.

These are quite impressive results given that some ties and inversions would be expected since (1) some paragraphs might actually be essentially as relevant to other paragraphs as to their own translations, (2) the LSI representation is obviously neither complete nor entirely accurate, and (3) the translations on which it is based are necessarily imperfect.

In the experiment described above, words that have identical spellings in the two languages are treated as a single term. This includes proper names and numbers as well as words that happen to have the same spelling in French and English. It is possible, though unlikely, that these cross-language homonyms alone are sufficient to allow documents to find their cross-language mates. The example document from Table 1 has four words that are shared by its French and English parts: "hon", "Erik", "Nielsen" and "situation," and perhaps words like these contribute significantly to the results.

To rule out the possibility that our initial results were due to cross-language homonyms, we performed two experiments. First, we replicated the mate-finding study using a comparable (same preprocessing, term weighting and similarity metrics) direct vector retrieval algorithm. That is, we matched the English test documents directly against the French test documents without using the CL-LSI analysis. This method, which is only sensitive to exact term matches between the two languages, performed significantly better than chance (which would be $1/1500 = 0.1\%$). However, the results, given in Table 3, demonstrate that word overlap alone is insufficient to perform good mate finding in this test collection; it does not account for the impressive performance of CL-LSI.

A second interesting question to ask is whether CL-LSI can function when there is no word overlap at all. To measure this, we prepared a version of the document collection in which words appearing in French documents were assigned the prefix "F" and words appearing in English documents were assigned the prefix "E". As a result of this preprocessing, every pair of French and English documents has zero words in common. We repeated the mate-finding experiment under these conditions and obtained results comparable to the initial results—perhaps slightly better (see Table 4). By construction, the vector method results in performance at the chance level (0.1%). This indicates that the CL-LSI method is able

query	document	returned mate at rank 1	percentage
English	French	1481	98.7%
French	English	1487	99.1%

Table 4: Results for cross-language mate finding using CL-LSI with no word overlap.

query	document	returned mate at rank 1	percentage
French translated to English	English	1,491	99.4%
English	French translated to English	1,489	99.3%
English translated to French	French	1,480	98.7%
French	English translated to French	1,486	99.1%

Table 5: Results for cross-language mate finding by MT-LSI.

to automatically find good language-independent representations, even when the languages involved have no words in common.

3.2.2 Cross-language Retrieval via Machine Translation

Although automated machine translation is far from perfect, it may be sufficient for the purpose of cross-language information retrieval. To test this hypothesis, we replicated the mate-finding experiment from the previous section using a method we call MT-LSI. First, we removed the French paragraphs from the 982-document training set and created a 982-dimensional English-only LSI space. We then folded in the 1,500 English-only test documents.

We next used a publicly available machine translation system from SYSTRAN [8, 13] to translate each of the 1,500 French-only test documents into English. These automatically translated documents were then folded in to the English-only space.

Now it is possible to treat each of the 1,500 English documents as queries, and to ask, if we automatically translated the 1,500 French documents into English and compared each query to the translated documents, how often would the (translated) cross-language mate be returned first? Similarly, we could ask what would happen if the French documents were translated and used as queries for the English documents or if we reversed the roles of French and English and performed the comparisons in a French-only space. Table 5 summaries the result of these experiments. In contrast to some earlier work [7, 1], we did not find that query translations results in large performance drops; we attribute this to the fact that our "queries" are document-length objects.

The results for the MT-LSI method were much stronger than we expected. The success rates were easily as good as those achieved by the CL-LSI method. Looking at the translations themselves, we noticed that they were readable, although they contained some awkward grammar. Table 6 contains the translation of the example French paragraph from Table 1. It is interesting to note that many vocabulary terms are not translated optimally for this domain: "Defense nationale" should be "National Defence," and "President" should be "Speaker" (see Table 1). Nevertheless, the combination of machine translation and the English-only (or French-only) semantic space is sufficient for excellent retrieval.

The hon. Erik Nielsen (Deputy Prime Minister and Minister for Defense nationale): Mr. President, we are in constant communication with our representatives consular in Libya. According to our information, the situation is stabilizing itself, and the Canadians are not immediately threatened. Consequently, my answers of yesterday, whose representative had to take note, did not change.

Table 6: Automatic English translation of a French document.

	English	French
CL-LSI	90.5%	55.4%
MT-LSI	91.7%	62.9%

Table 7: Results for top 1 English pseudo-query retrieval.

It is also worth considering how well a machine translation system without any LSI analysis would perform. To assess this, we repeated the experiment using the ordinary vector method for query-document comparisons (MT-vector) and found results almost identical to MT-LSI for both the mate retrieval task and the pseudo-query task described in the next section. Since the MT-vector and MT-LSI algorithms performed so similarly, we only report the MT-LSI results in this paper.

In spite of the fact that MT-LSI achieved comparable results to the CL-LSI method, we feel that the CL-LSI method still possesses certain advantages. Creating a CL-LSI system for a new document collection is substantially easier than creating a new machine-translation program. The skills required for a human to create the dual-language documents needed for training are more common than the skills required to build a software system as complex as a machine translator. The fact that the CL-LSI system performs comparably to a highly developed MT program is strong support for the claim that CL-LSI is practical, accurate and cheap.

3.2.3 Automatically Generated Short Queries

In the cross-language mate-finding experiments described above, queries are as long as documents, a situation that is known to result in good retrieval. To simulate the more realistic scenario in which user queries are a good deal shorter, we used the English-only LSI space to create "pseudo-queries."

We created pseudo-queries for each of the 1,500 test documents by finding the 5 nearest terms in the English-only space to each English test document. For the English half of the example document in Table 1, the generated pseudo-query was "consular immediate inundated threat nielsen."

We used these pseudo-queries to find the top 10 English documents and the top 10 French documents using the CL-LSI and MT-LSI methods and measured the percentage of times that the document corresponding to the pseudo-query was returned as the best match. Results are tabulated in Table 7. (The MT-LSI/French entry corresponds to matching the English pseudo-query to the automatic translation of the French test documents.)

	English	French
CL-LSI	99.8%	92.3%
MT-LSI	99.9%	92.0%

Table 8: Results for top 10 English pseudo-query retrieval.

The column labeled "English" records the percentage of pseudo-queries that correctly select out their corresponding document (i.e., the document used to generate them in the first place). We see that these pseudo-queries are not perfectly accurate, but are fairly good descriptions of the documents they are intended to match.

The column labeled "French" shows that the MT-LSI and CL-LSI systems were successful in matching the short pseudo-queries to the corresponding French documents in only about 6 out of 10 of the cases. A close look at the raw data indicated that both systems were assigning the relevant French document relatively high ranks and were assigning highly similar documents the first rank.

To make the measure a little less sensitive to the additional noise introduced by shorter queries, we also measured the percentage of times the pseudo-queries returned their target documents anywhere in a list of the top 10 matches. These results are summarized in Table 8; they compare quite favorably to the results with full-length queries described earlier, at the expense of slightly lower precision.

3.2.4 Human-Generated Short Queries and Relevance Judgments

While the results of the previous section are quite impressive for such short queries, they are still less informative than good human-generated queries and human relevance judgments. We have collected a set of 4 queries in English from each of 8 subjects that can write in English and read in French. We are currently in the process of gathering relevance judgments for a set of French documents. We will report on these results in a later paper.

4 Conclusion

We sketched an approach called CL-LSI for cross-language retrieval using LSI along with some of the simple tests we have run to measure its usefulness. Other researchers have experimented with the CL-LSI method using other test collections and other languages and have obtained positive results: Berry and Young [2] used Greek and English versions of the Gospel, Oard [11] used Spanish and English documents in a text filtering task, Landauer, Littman and Stornetta [10] used English and Japanese abstracts of scientific papers. Important questions remain concerning the number and size of documents to be used during training, and it is important that standardized collections be created to allow precision-recall tests.

In conclusion, by all available measures, the CL-LSI system works quite well. It automatically finds a language-independent representation for documents that is sufficient to identify relevant documents in one language using long and short queries in another language. We

also examined a method we called MT-LSI that uses automatic machine translation to make cross-language comparisons. This method also performed quite well in our tests.

References

[1] Lisa Ballesteros and Bruce Croft. Dictionary methods for cross-linguistic information retrieval. Presented at SIGIR'96 workshop on cross-linguistic information retrieval, 1996.

[2] M. W. Berry and P. G. Young. Using Latent Semantic Indexing for multilanguage information retrieval. *Computers and the Humanities*, 29(6):413–429, 1995.

[3] J. K. Cullum and R. A Willoughby. Chapter 5: Real rectangular matrices. In *Lancozos algorithms for large symmetric eigenvalue computations - Vol 1 Theory*. Brikhauser, Boston, 1985.

[4] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[5] S. T. Dumais. Using LSI for information filtering: TREC-3 experiments. In D. Harman, editor, *The Third Text Retrieval Conference (TREC3)*, pages 219–230. National Institute of Standards and Technology Special Publication 500-225, 1995.

[6] G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In *Proceedings of the 11th ACM International Conference on Research and Development in Information Retrieval*, pages 465–480, 1988.

[7] David A. Hull and Gregory Grefenstette. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–57. Association for Computing Machinery, 1996.

[8] William John Hutchins and Harold L. Somers. *An Introduction To Machine Translation*. Academic Press, San Diego, 1992.

[9] Thomas K. Landauer and Michael L. Littman. Fully automatic cross-language document retrieval using latent semantic indexing. In *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, pages 31–38. UW Centre for the New OED and Text Research, Waterloo Ontario, October 1990.

[10] Thomas K. Landauer, Michael L. Littman, and Wakefield S. Stornetta. A statistical method for cross-language information retrieval. Unpublished manuscript, 1992.

[11] Douglas William Oard. *Adaptive Vector Space Text Filtering for Monolingual and Cross-Language Applications.* PhD thesis, University of Maryland, College Park, August 1996.

[12] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

[13] SYSTRAN. Systran software html translation page! Available through URL http://www.systranmt.com/translate.html, 1996.