Automatic Cross-Language Retrieval Using Latent Semantic Indexing

Susan T. Dumais
Todd A. Letsche
Info. Sciences Research
Bellcore
Morristown, NJ 07960
std | letsche@bellcore.com

Michael L. Littman
Dept. of Computer Science
Duke University
Durham, NC 27708-0129
mlittman@cs.duke.edu

Thomas K. Landauer
Dept. of Psychology
University of Colorado
Boulder, CO 80309-0344
landauer@psych.colorado.edu

Abstract

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 multilingual semantic space using Latent Semantic Indexing (LSI). Strong test results for the cross-language LSI (CL-LSI) method are presented for a new French-English collection. We also provide evidence that this automatic method performs comparably to a retrieval method based on machine translation (MT-LSI), and explore several practical training methods. By all available measures, CL-LSI performs quite well and is widely applicable.

Introduction

Cross-language LSI (CL-LSI) is a fully automatic method for 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).

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 also examine several practical training issues.

Overview of Latent Semantic Indexing (LSI)

Most information retrieval methods depend on exact matches between words in users' queries and words in documents. 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 in which words that occur in similar contexts are near each other. 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 (Cullum and Willoughby, 1985). 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 (Salton and McGill, 1983). This term-document matrix is decomposed into a set of k, typically

Appears in: **AAAI-97 Spring Symposium Series: Cross-Language Text and Speech Retrieval.** March 24-26, 1997, Stanford University, pp. 18-24.

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 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. Doc 3 contains term 2, Doc 1 contains term 1, and Doc 2 contains both, but the terms are uncorrelated. In contrast, LSI represents terms as continuous values on each of the orthogonal indexing dimensions. Terms are not 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. Deerwester et al. (1990) and Furnas et al. (1988) present detailed mathematical descriptions and examples of the underlying LSI/SVD method.

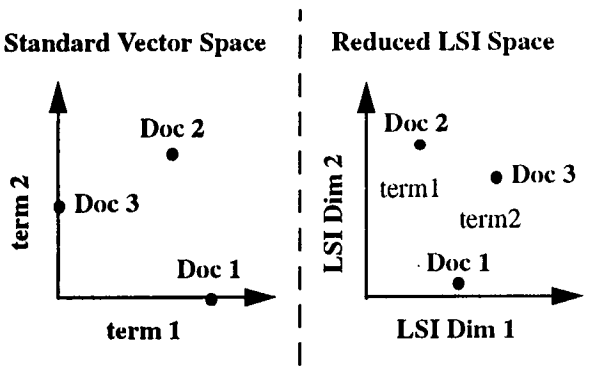

Figure 1. Term representations in the standard vector vs. reduced LSI vector models.

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 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. A document is located at the weighted vector sum of its constituent terms. 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 (Deerwester et al., 1990; Dumais, 1995).

Cross-Language Retrieval Using LSI

Landauer and Littman (1990) first described how LSI could easily be adapted to cross-language retrieval. 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) is given in Table 1.

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

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.

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. Because the training documents contain both French and English terms, the LSI space will contain terms from both languages (term1 through term3 in English and *mot1* through

mot4 in French), and the training documents (**EFDoc**). This is what makes it possible for the CL-LSI method to avoid query or document translation. Words that are consistently paired (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.

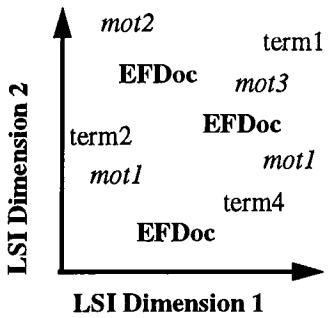

Figure 2. The training phase of CL-LSI. Training items are dual English-French documents, and words from both languages are located in the CL-LSI space.

The next step in the CL-LSI method is to add (or "fold in") documents in just French or 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** or ***FDoc***). 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 (dashed vector) or English (solid vector) and get back the most similar documents regardless of language.

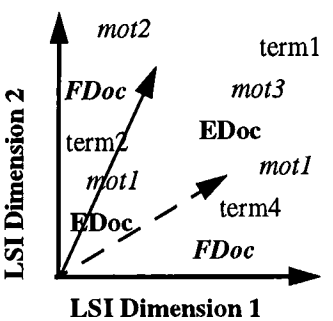

Figure 3. The fold-in and query phases of CL-LSI. Monolingual documents are located at the vector sum of their constituent terms.

Experimental Tests

Landauer and Littman (1990) describe the first 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, Littman, Dumais and Landauer (1997) replicated and extended these results. They 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. The 982 dual-language documents were used to create a dual-language semantic space. 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. (On a standard Sparc workstation, this type of analysis takes about 2 minutes.) We describe these results in some detail because they provide good background and baselines for the new training results we report in this paper.

Cross-language Mate Retrieval.

Since no standard multi-language test collection with cross-language queries and relevance judgments was available to evaluate the CL-LSI retrieval system, they used what we call a mate retrieval test. This test uses documents to find their cross-language mates and can be thought of as treating each of the 1,500 English documents as queries, each with exactly one relevant document in French---its translation (or mate). (The same test is conducted with French documents as queries and English documents as targets.) The results are presented in the first row of Table 2, which show that the CL-LSI method does an excellent job of retrieving cross-language mates first.

TABLE 2. Cross-language mate finding using CL-LSI and a no-LSI control. Percent of cross-language mates retrieved first.

	Eng->Fr	Fr->Eng	Average
CL-LSI	98.3%	98.5%	98.4%
no-LSI	47.7%	49.5%	48.6%

These are quite impressive results given that some paragraphs might actually be essentially as relevant to other para-

graphs as to their own translations, and the translations on which it is based are necessarily imperfect.

It is possible, though unlikely, that 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.

Littman et al. (1997) replicated the mate-finding study using the standard vector method without any LSI dimension reduction. This method (no-LSI), which is only sensitive to exact term matches between the two languages, performed significantly worse than CL-LSI as shown in the second row of Table 2. Word overlap alone is insufficient to account for the impressive performance of CL-LSI.

A related and very important question is whether CL-LSI can function when there is no word overlap at all. To measure this, they 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. They repeated the experiment under these conditions and obtained results comparable to the initial results---perhaps slightly better (first row Table 3). By construction, the vector method results in performance at the chance level (0.1%). This indicates that the CL-LSI method is able to automatically find good language-independent representations, even when the languages involved have no words in common.

TABLE 3. Cross-language mate finding using CL-LSI and a no-LSI control, when languages have no word overlap. Percent of cross-language mates retrieved first.

	Eng->Fr	Fr->Eng	Average
CL-LSI	98.7%	99.1%	98.9%
no-LSI	.1%	.1%	.1%

Cross-language Retrieval w/ Machine Trans.

Although automated machine translation is far from perfect (see Table 4 which contains the automatic translation of the French paragraph from Table 1), it may be sufficient for the purpose of cross-language information retrieval. To test this Littman, et al. (1997) replicated the mate-finding experiment using machine translation (MT).

TABLE 4. Machine translation of French section of Table 1.

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.

First, they created a 982-dimensional English-only LSI space and folded in the 1,500 English-only test documents. They then used a publicly available machine translation system (Hutchins and Somers, 1992; Systran, 1996) to translate the 1,500 French-only test documents into English. These automatically translated documents were then folded in to the English-only space. Table 5 summarizes the result of these experiments. In contrast to some earlier work (Hull and Grefenstette 96; Ballesteros and Croft 96), they did not find that query translations resulted in large performance drops, and attribute this to the fact that their "queries" were document-length objects. Results were essentially the same for the ordinary vector method without any LSI.

TABLE 5. Cross-language mate finding using MT-LSI.

query	document	percent
French translated to Eng	English	99.4%
English	French translated to Eng	99.3%
English translated to Fr	French	98.7%
French	English translated to Fr	99.1%

Short Queries.

To simulate more realistic retrieval scenarios in which user queries are much shorter, they created English "pseudo-queries" by finding the 5 nearest terms to each English test document. The pseudo-query generated for the English part of Table 1 was "consular immediate inundated threat nielsen". They used these pseudo-queries to find the top 1 or 10

French documents using the CL-LSI and MT-LSI methods. Results are shown in Table 6.

TABLE 6. English pseudo-query retrieval.

	Top 1 EngP->Fr	Top 10 EngP->Fr
CL-LSI	55.4%	92.3%
MT-LSI	62.9%	92.0%

Both methods were successful in matching the short pseudo-queries to the corresponding French documents in only about 6 out of 10 of the cases. The results for the top 10 compare quite favorably to the results with full-length queries described earlier, at the expense of slightly lower precision.

We have extended these results by using humanly generated short queries. We were able to obtain English and French versions of Yellow Page category labels. Examples are shown in Table 7. The Yellow Page headings average 2.5 words in length in both English and French.

TABLE 7. English and French Yellow Pages category labels.

English	French
banks	banques
cleaners	nettoyeurs
disc jockeys	discothèques mobiles
monuments	monuments
sun tan salons	salons bronzage

We can now use these natural short queries to retrieve their cross-language mates in a CL-LSI space. We had a CL-LSI space available from a small training of the Hansard corpus, but the domains and vocabularies are obviously quite different. While the Canadian Parliament proceedings often cover tax law reform or wheat prices, there was no mention of tanning salons or flat tires in our sample. A CL-LSI space was created using these documents and 145 English and French Yellow Page category labels were folded in. The results for 145 queries are shown in the first row of Table 8. These results are the average performance for English queries retrieving French categories and vice versa. While performance is far from perfect, it is also a good deal better than chance (.7%), suggesting that important cross-language relationships are being represented.

We were unable to obtain more general and appropriate parallel (or comparable) collections in French and English for training purposes, so we explored some alternative training methods. We chose to take advantage of a commercially available machine translation system to translate English corpora into French and create dual-language training documents. Obviously in cases where no commercial tools existed, one would have humans generate translations of a small number of training documents, but this was impractical given our resources.

We first compared performance using the original Hansard corpus which has human-generated parallel texts for training with a CL-LSI space constructed using machine translation. To do this we used Systran to translate the 2482 English paragraphs from our Hansard sample into French. We created dual-language documents using the English and machine-translated French pairs and derived a 330 dimension CL-LSI. The 145 English and French Yellow Page category labels were folded in and tested as described above. The results are shown in the second row of Table 8 (Hansard corpus, machine translation to generate dual-language training documents). Performance is about 10% worse for retrieving the corresponding category first, but 15% better when looking at the top 10.

It is important to note that machine translation is used only to create the dual-language training documents. Subsequently CL-LSI, which does not involve any translation, is used for retrieval tests.

TABLE 8. Yellow Page cross-language retrieval using CL-LSI under several training conditions. Collections marked with an * indicate that machine translation was used to create the dual-language documents for training.

	training size	Top 1	Top 10
Hansard	2482	22.8%	47.2%
Hansard*	2482	20.0%	54.3%
Encyclopedia*	30473	52.4%	80.3%
YP-www*	3515	63.8%	86.9%

So, useful relationships can be derived from imperfect machine translations. For present purposes this is important for training when parallel dual-language corpora are not available. This enables us to construct more generally useful CL-LSI representations for cross-language retrieval. We looked at a CL-LSI space based on an online Encyclopedia containing 30,473 articles. Results are shown in the third row of Table 8, and are a good deal better than those obtained with the restricted Hansard CL-LSI space. Performance is only slightly worse that seen in Table 6 for the Hansard pseudo-queries in a Hansard space; and the Hansard queries were twice as long. This suggests that a general purpose CL-LSI space will be quite useful for a variety of cross-language retrieval applications.

We also looked at a new corpus we created to contain texts more closely related to Yellow Page categories. This YP-www corpus was created using 110 different Yellow Page categories as queries to popular WWW Search Services. We submitted each of these 110 categories to 5 Search Services (Alta Vista, Excite, HotBot, InfoSeek, Lycos), retrieved the 10 best-matching URLs, and then fetched the full text associated with them. The resulting 3515 items were translated by machine to construct dual-language training documents. Cross-language retrieval performance using the CL-LSI space derived from this collection is also quite good as seen in the last row of Table 8. Many fewer training documents were used here, but the coverage and vocabulary is more closely related to the test items.

We suspect we could do somewhat better in tuning the training collection to the retrieval application. First, the categories we used as seeds were different than those we used for testing resulting in less than perfect overlap in vocabularies. In addition, the text associated with the URLs was far from ideal. Retrieved documents sometimes contained long lists of all the Yellow Page categories, and such documents do not provide very good context to define the inter-relationships among words. Finally, we translated the texts by machine. In spite of these less than optimal training conditions, performance in quite reasonable. The costs of building a training collection in this way are really quite small. The average English document was only 300 words long. If it took a human 15 minutes to translate each training document, developing the training collection (and the resulting fully automatic cross-language retrieval system) would take only 875 person hours. This is substantially less time than is invested in developing machine translation systems for new pairs of languages.

We also compared the CL-LSI results obtained for this new test collection with two control conditions. First, we looked at the standard vector method without any translation (no-LSI). Success here depends on the degree to which the Yellow Page categories share important words in the two languages. Second, we looked at performance using machine translation of the queries (MT). Results of these new tests are shown in Table 9.

TABLE 9. Yellow Page cross-language retrieval for YP-www comparing CL-LSI with control conditions.

	Top 1	Top 10
CL-LSI	63.8%	86.9%
no-LSI	15.1%	28.9%
MT	57.5%	74.8%

As Littman et al. (1997) found, word overlap alone is clearly insufficient to account for the success of the CL-LSI approach. In addition, CL-LSI is about 15% more accurate than machine translation in this test. We suspect this is because the queries are very short with little room for translation errors.

Conclusions

We sketched an approach called CL-LSI for cross-language retrieval using LSI, and described several tests of its usefulness. Although we have reported results for only French-English collections, other researchers have experimented with the CL-LSI method using other test collections and other languages and have obtained equally positive results. Berry and Young (1995) used Greek and English versions of the Gospel; Oard (1996) used Spanish and English documents in a text filtering task; and Landauer, Littman and Stornetta (1992) used English and Japanese abstracts of scientific papers. In addition, Frederking et al. (1997) have reported success with statistically-based dimension reduction approaches (both LSI and the generalized vector space model) to cross-language retrieval.

By all available measures, the CL-LSI system works very 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. CL-LSI produces results comparable to (and sometimes better than) those obtained with well-tuned machine translation systems at substantially less cost. Creating a CL-LSI system for a new document collection is s much 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.

References

Ballesteros, L. and Croft, B. Dictionary methods for cross-linguistic information retrieval. Presented at SIGIR'96 Workshop on Cross-linguistic Information Retrieval, 1996.

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

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

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. A. Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41(6)*, 391-407, 1990. [http://superbook.bellcore.com/~std/papers/JASIS90.ps]

Dumais, S. T. Using LSI for information filtering: TREC-3 experiments. In D. Harman (Ed.) *The Third Text Retrieval Conference (TREC3), National Institute of Standards and Technology Special Publication 500-225*, 219-230, 1995. [http://superbook.bellcore.com/~std/papers/TREC3.ps]

Frederking, R., Mitamura, T., Nyberg, E., and Carbonell, J. Translingual information retrieval. Paper presented at AAAI-97 Spring Symposium Series, Cross-Language Text and Speech Retrieval.

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

Hull, D. A. and Grefenstette, G. Querying Across Languages: A Dictionary-based approach to multilingual information retrieval. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 49-57, 1996.

Hutchins, W. J. and Somers, H. L. *An Introduction To Machine Translation.* San Diego, Academic Press, 1992.

Landauer, T. K. and Littman, M. L. Fully automatic cross-language document retrieval using latent semantic indexing. *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research.* UW Centre for the New OED and Text Research, Waterloo Ontario, 31-38, October 1990. [http://www.cs.duke.edu/~mlittman/docs/x-lang.ps]

Landauer, T. K., Littman, M. L. and Stornetta, W. S. A statistical method for cross-language information retrieval. Unpublished manuscript, 1992.

Littman, M. L., Dumais, S. T. and Landauer, T. K. Automatic cross-linguistic information retrieval using Latent Semantic Indexing. To appear in G. Gredenstette (Ed.), *Cross Language Information Retrieval*, 1997. [http://superbook.bellcore.com/~std/papers/XLANG96.ps]

Oard, D. W. Adaptive vector space text filtering for monolingual and cross-language applications. Ph.D. Thesis University of Maryland, College Park, 1996. [http://www.ee.umd.edu/medlab/filter/papers/thesis.ps.gz]

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

SYSTRAN. SYSTRAN Software HTML Translation Page, 1996. [http://www.systranmt.com/translate.html]