Automatic 3-Language Cross-Language Information Retrieval with Latent Semantic Indexing

Bob Rehder
Dept. of Psychology
Inst. of Cognitive Science
U. of Colorado, Boulder
Boulder, CO 80309
rehder@psych.colorado.edu

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

Susan Dumais
Microsoft Research
One Microsoft Way
Redmond WA, 98052
sdumais@microsoft.com

Thomas K. Landauer
Dept. of Psychology
Inst. of Cognitive Science
U. of Colorado, Boulder
Boulder, CO 80309
landauer@psych.colorado.edu

Abstract

This paper describes cross-language information-retrieval experiments carried out for TREC-6. Our retrieval method, cross-language latent semantic indexing (CL-LSI), is completely automatic and we were able to use it to create a 3-way English-French-German IR system. This study extends our previous work in terms of the large size of training and testing corpora, the use of low-quality training data, the evaluation using relevance judgments, and the number of languages analyzed.

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); this semantic space is exploited in the form of a *vector lexicon*, which assigns each word in each language to a point in the high-dimensional space.

For the CL-LSI method to be used, an initial sample of documents in one language must be available with "mates" in all other languages. In past work, these mates were created by human translators; in the present work, we used a combination of machine translation and automatic mate selection from a comparable corpus to create mate sets. An LSI analysis of the set of documents and mates results in a multi-language *semantic space* in which terms from all languages are represented. Concretely, this semantic space takes the form of a vector lexicon in which each word in each of the languages is assigned a high-dimensional vector representation. Queries in any language can retrieve documents in any language without the need to translate the query because all text records (documents and queries) are represented as language-independent numerical vectors in the same semantic space.

The present work builds on our past experience with CL-LSI by

1. scaling to larger document collections than had been previously attempted,
2. using much noiser training data (no human translations) than had been previously attempted, and
3. using more languages than had been previously attempted (3 instead of 2).

We explored a completely automatic approach to information retrieval between topics in English, French, and German and documents in English, French, and German. To train our system, we began with a coarsely parallel aligned collection of over 80,000 German and French documents provided by NIST, which we used to train an initial German-French cross-language retrieval system. The German-French training pairs that were assigned the lowest similarity scores by our initial system were discarded in an attempt to weed out document pairs that were not properly aligned. The remaining 40,000 French-German pairs were then augmented with computer-generated English translations of the German documents (also provided by NIST), to create a collection of 40,000 3-language mate triplets, which we used to train an English-French-German retrieval system. Thus, in contrast to most previous work on automatic cross-language IR, no human translations were used in training. We used our retrieval system to compare both short and long queries in all three languages against the full set of English, French, and German documents. We feel that, by analogy to our experience with monolingual LSI, our approach is likely to show the largest benefits for the short topics, but this has been difficult to assess so far. The noteworthy aspects of our approach are that it is completely automatic, it works between any pair of the three target languages, it is trained using an imperfectly aligned collection, and it exploited a simple "bootstripping" technique to help make the most of noisy training materials.

Background

This section provides background on latent semantic indexing (LSI) and its cross-language extension. Other introductions are also available (Deerwester *et al.* 1990; Berry, Dumais, & O'Brien 1995; Dumais 1995).

Latent Semantic Indexing Motivation

Latent semantic indexing is a variant of the vector-space method (Salton & McGill 1983) 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 (Cullum & Willoughby 1985). For information retrieval and filtering applications we begin with a large term-document matrix, in much the same way as vector-space or Boolean methods do. This term-document matrix is decomposed into a set of k, typically 200–300 in monolingual applications, 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.

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.

In monolingual document-retrieval tests, 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).

Latent Semantic Indexing Mathematics

LSI begins with a collection of m documents containing n unique terms and forms an $n \times m$ sparse matrix E, with E_{ij} containing a value related to the number of times term i appears in document j. Various weighting schemes can be applied to the raw occurrence counts; in this work, we used log-entropy weighting ($\log(\mathrm{tf}+1)$entropy).

Once the document-term matrix E has been created, LSI computes the similarity between two text objects (a query and a document, say) as follows. First, a text object q is represented by an $n \times 1$ vector, much like a column of the E matrix and with the same sorts of term weighting applied. Next, the similarity between text objects q_1 and q_2 can be computed, typically by cosine scoring; in the vector-space method, this can be represented as $\mathrm{sim}(q_1, q_2) = q_1^T q_2 / \sqrt{q_1^T q_1 \cdot q_2^T q_2}$.

A mathematically useful way of viewing the process of computing text-object similarity scores in the vector-space method is this. Each of the n terms in the collection has a vector representation, specifically term i is an $n \times 1$ vector of zeros with a 1 in component i. The representation of a text object q is a weighted sum of the term vectors of the terms that appear in the text object. Thus, the similarity between text objects q_1 and q_2 is

$$\mathrm{sim}(I_n q_1, I_n q_2), \tag{1}$$

where I_n is the $n \times n$ identity matrix. Here, I_n plays the role of a *vector lexicon*, in that it assigns each term a vector "definition." Of course, premultiplying by the identity matrix in Equation 1 does not change the comparison in any way; by using other vector lexicons, we can substantially change the way similarities are computed. Note that the only role played by the document-term training matrix E in the vector-space method is in the computation of weighting factors for the components of text objects.

LSI can be viewed very similarly to the vector-space method. LSI also begins with the formation of the term-document matrix E. Then, the E matrix is analyzed using singular value decomposition (SVD) to extract structure concerning document-document and term-term correlations. Mathematically, an SVD of E can be written

$$E = U(E)\ \Sigma(E)\ V(E)^T, \tag{2}$$

where $U(E)$ is an $n \times n$ matrix such that $U(E)^T U(E) = I_n$, $\Sigma(E)$ is an $n \times n$ diagonal matrix of *singular values* and $V(E)$ is an $n \times m$ matrix such that $V(E)^T V(E) = I_m$. This assumes for simplicity of exposition that E has fewer terms than documents, $n < m$.

This SVD analysis can be used to construct lower rank approximations of E, and this is how it is typically used in the context of LSI. Reducing the rank of the approximation results in a synonym-collapsing effect in practice. It also reduces the total amount of processing and storage associated with preprocessing and retrieval. We write

$$E_k = U_k(E)\ \Sigma_k(E)\ V_k(E)^T, \tag{3}$$

to denote the components of the k-dimensional SVD and its rank-k reconstruction of E.

The $U_k(E)$ matrix in Equation 3 can be used as an alternative vector lexicon to the I_n in Equation 1 in that it assigns a vector representation to every term in the term-document matrix E. Thus, in LSI, the k-dimensional similarity between text object q_1 and text object q_2 in the context of E is

$$\mathrm{sim}(U_k(E)^T q_1, U_k(E)^T q_2). \tag{4}$$

Berry, Dumais, & O'Brien (1995) give justifications for the use of the matrix of left singular vectors $U_k(E)$ as a vector lexicon.

Cross-language LSI

The techniques of mono-lingual LSI transfer easily to the cross-language case simply by using a different notion of the term-document matrix (Landauer & Littman 1990).

For concreteness, let E be a term-document matrix of m English documents and n^E English terms, F be a term-document matrix of m semantically equivalent French documents and n^F French terms, and G be a term-document matrix of m semantically equivalent German documents and n^G French terms. These documents are mate-aligned, in the sense that document $1 \leq i \leq m$ in the English collection is directly related to document i in the French and German collections. The multi-language term-document matrix

$$M = \begin{bmatrix} E \\ F \\ G \end{bmatrix}$$

is an $(n^E + n^F + n^G) \times m$ matrix in which column i is a vector representing the English, French, and German terms appearing in the union of document i expressed in all three languages.

Cross-language LSI (CL-LSI) begins with the matrix M and performs an SVD,

$$M = \begin{bmatrix} U_k^E(M) \\ U_k^F(M) \\ U_k^G(M) \end{bmatrix} \Sigma_k(M)\ V_k(M),$$

where $U_k^E(M)$, $U_k^F(M)$, $U_k^G(M)$ are k-dimensional vector lexicons for English, French, and German, respectively. Empirically, similar English, French, and German words are given similar definitions, so this vector lexicon can be used for cross-language retrieval. In particular, consider an English text object q_E and a French text object q_F. They can be compared using the obvious generalization of Equation 4,

$$\mathrm{sim}(U_k^E(M)^T q_E, U_k^F(M)^T q_F). \tag{5}$$

Note that, in our experiments, we chose not to take advantage of cross-language homonyms. That is, the word "documents" in French was treated distinctly from the word "documents" in English. The same holds true of names and numbers. For this collection of languages, it is likely that identifying and exploiting cross-language homonyms would improve performance. We chose not to do this so that we could better evaluate how well CL-LSI was able to identify patterns in word usage between the languages without relying on cognates or other "incidental" properties of the languages used in this study.

Previous Evaluations of CL-LSI

In past work, we have used a number of informal evaluation techniques to help determine whether retrieval systems created using CL-LSI can effectively compare text objects between languages. In the *overlap* technique (Landauer & Littman 1990), we began with a set of several thousand English and French mates and compared each with a set of English queries to determine the 10 best matching French documents and 10 best matching English documents to each query. We then counted, for each query, the number of mates in common in the English and French return sets, and found that an average of 4.1 mate pairs appeared in the return sets. Thus, to the extent that CL-LSI is able to match English queries to English documents, it is also able to match English queries to French documents nearly as well.

In *mate-retrieval* evaluation, we again begin with a test set of English and French mates. Next, we take each English document as a query and compute the rank of its French mate when the English "query" is compared with each French document. While this use of long queries does not provide a very accurate measure of the performance of CL-LSI in a real retrieval setting, it does give some indication as to whether the language-independent vector representation of meaning is at all reasonable. A typical result (Dumais, Landauer, & Littman 1996) is that CL-LSI returns cross-language mates over 98% of the time from a test set of 1,500 mates. Less strong, but still impressive, results are obtained using imperfectly matched or machine translated mates for training, and mismatches between training and testing data (Dumais, Littman, & Landauer 1997 to appear).

CL-LSI has been evaluated in a more traditional *relevance-judgment* experiment (Carbonell *et al.* 1997). The implementation of CL-LSI in that study was compared to the generalized vector-space method, example-based query translation, and pseudo-relevance feedback query expansion, as well as a number of other techniques, and fared relatively poorly. The version of CL-LSI in our work differs from that in the Carbonell *et al.* (1997) study in our use of Equation 5 (instead of $\mathrm{sim}(U_k^E(M)^T\Sigma_k(M)^{-1}q_E, U_k^F(M)^T\Sigma_k(M)^{-1}q_F))$ and in the number of dimensions used (we tend to use 500-1500 dimensions for cross-language comparisons, they used 200).

In our recent studies with the Carbonell *et al.* (1997) collection, CL-LSI outperformas all methods except example-based query translation.

CL-LSI in TREC-6

The document collection in the TREC-6 cross-language track experiments consisted of English, German, and French newspaper articles from 1988-1990. Specifically:

- English (242,918 documents, 684MB):
  Associated Press newswire (AP)
- German (185,099 documents; 269MB):
  Schweizerischen Depeschenagentur, Swiss news agency (SDA-G)
- German (66,741 documents, 176MB):
  Neue Zuercher Zeitung, Swiss German newspaper (NZZ)
- French (141,656 documents; 199MB):
  Schweizerischen Depeschenagentur, Swiss news agency (SDA-F).

A total of 25 topics were created by NIST in each of English, French, and German. We used an initial alignment, bootstrip cleaning, and machine translation to create an English-French-German CL-LSI system. Each of these steps are explained in more detail in the following sections.

Initial Alignment

For CL-LSI to apply in English, French, and German, it is necessary to have a set of documents in one language with mates in both of the others. In previous studies, corpora were used in which human translators had generated these mates. In this study, we used automatic methods to generate these mates. Note that neither the alignment (Sheridan & Ballerini 1996), nor the machine translation was actually carried out by our group. We simply used the results of the work of others.

To begin, the SDA-G set (185k German documents) was taken as the core set of training documents. For each of these, the SDA-F set (142k French documents) was searched for possible mates. As truly accurate mate decisions would require human judgment, a simple rule of thumb was used. A French document was declared as a mate for a German document if the two documents (newswire articles) appeared on the same day and had a sufficient number of words in common in their keyword fields. Any German document with no mates found through this procedure was removed from the core set of training documents. The result was a set of 83,698 German documents and their automatically discovered French mates.

Note that it is possible for a single French document to occur more than once as a mate if it happened to align well with more than one German document. It is estimated that 20% of the aligned German documents are paired with a non-unique French document (a French document that is aligned with some other German one). Also, in our preliminary look at these alignments, we coarsely estimate that 10% of the 84k document pairs are misaligned. For example, one pair consists of articles about a bus bombing and a revenge killing in Jerusalem on the same day. These stories are not really related, but this is impossible to tell based on the keywords (terrorism, Jerusalem), or the date.

Bootstrip Cleaning

As an attempt to automatically weed out this sort of error, we carried out the following "bootstripping" procedure.

1. Separate the 83,698 French-German document pairs into a verification set of 3k documents and a training set of 80,698 documents.
2. Use CL-LSI on the 80,698 training documents to create a vector lexicon for German and French. As a check, calculate mate-retrieval scores for the

Training-set size	˜80k	40k	20k
Dimensions	500	1000	1200
Average cosine of mates	0.508	0.469	0.459
Average rank of mate	3.00	2.51	2.73
Average log rank	0.287	0.274	0.325
Proportion in top one	80.1%	82.5%	79.8%
Proportion in top ten	96.7%	96.6%	95.6%
Proportion in top fifty	99.6%	99.4%	99.4%
Number not in top 5%	4	1	1

Table 1: Mate-retrieval results for bootstripping experiment

Training set	F-G	E-F-G	E-F-G
Testing set	G-F	G-F	E-F
Dimensions	1000	800	800
Average rank of mate	2.51	2.76	2.97
Average log rank	0.274	0.308	0.284
Proportion in top one	82.5%	80.5%	82.2%
Proportion in top ten	96.6%	96.1%	96.0%
Proportion in top fifty	99.4%	99.4%	99.5%
Number not in top 5%	1	2	4

Table 2: Three-language mate-retrieval scores

3k verification documents using the vector lexicon (see Table 1).

3. Calculate similarity scores for each of the 80,698 training documents using the derived vector lexicon. Identify the 40k pairs with the highest similarity scores.
4. Use CL-LSI on the best 40k training documents to create a new vector lexicon for German and French. As a check, calculate mate-retrieval scores for the 3k verification documents using the new vector lexicon.
5. Calculate similarity scores for each of the 40k training documents using the derived vector lexicon. Identify the 20k pairs with the highest similarity scores.
6. Use CL-LSI on the best 20k training documents to create a new vector lexicon for German and French. As a check, calculate mate-retrieval scores for the 3k verification documents using the new vector lexicon.

Looking at the average rank of mate reported in Table 1, we see that performance improves in going from 80k training documents to 40k, then degrades at 20k. Most of the other scores follow a similar trend. Next, we give a brief explanation of the measures given in Table 1. The average pairwise cosine gives the similarity between the mates in the 3k-document verification set. Log ranks were computed simply to diminish the weight given to pairs with very poor rankings (surpress the effect of outliers).

The best of these three runs appears to be the one with the 40k-document training set. It is this set we use in the remainder of our experiments. Mate-retrieval performance is quite strong for this training set: 82% of the mates have the highest cosine, 97% of mates are in the top 10, and over 99% are in the top 50. While there are obviously some pairs that are assigned poor similarity scores, there is only 1 document pair (out of 3k) that had a rank not in the top 5%, which is reasonable.

Machine Translation: Extending to 3 Languages

After the bootstrip cleaning, we were left with an French-German retrieval system. To extend this to include English, we made use of machine-translations of the SDA-G documents. Of the 40k German documents in our core training set, 12 of them did not have available translations into English. Therefore, we carried out a CL-LSI analysis of a 39,988-document collection of German, with a French and English mate for each German document. (Note that preliminary experiments (Littman & Keim 1997) indicate that it is important for 3-language training collections to use complete sets of mates.)

Table 2 gives mate-retrieval results for the resulting 800-dimensional three-way retrieval system. The German-to-French mate-retrieval performance degrades slightly for the 3-way system compared to the French-German system from the bootstapping experiment; it is not clear whether this is due to the inclusion of English or the decreased number of dimensions. Nevertheless, the 3-way system exhibits respectable English-to-French mate-retrieval performance—comparable to the German-to-French performance. This is remarkable, given that the statistical relationship between English and French in this system is very indirect: French document i was paired with English document i only because of the coarse alignment between German i and French i and the fact that German i was machine translated to yield English i.

Results

A noteworthy feature of the CL-LSI approach is that, because all words from all languages exist together in a common high-dimensional space, queries can be matched against documents in all languages together; without specifying the target language, a user receives first a document in whichever language gives the best match. While this feature was not exploited in the TREC studies, here is a simple illustration.

Figure 1 gives the English version of one of the topics. We issued a long query (the union of the

short: Reasons for controversy surrounding Waldheim's World War II actions.
long: Revelations about Austrian President Kurt Waldheim's participation in Nazi crimes during World War II are argued on both sides. Relevant documents are those that express doubts about the truth of these revelations. Documents that just discuss the affair are not relevant.

Figure 1: Long and short forms of topic 1

long and short version of the topic) against the entire test collection (all three languages). The top ten documents (and the 50th) retrieved, along with their language and similiarity score, are:

1. (German: 0.538): Waldheim als "Mitwisser," nicht als "Mitschuldiger."
2. (English: 0.532) Waldheim Says Pope Visit Will Help Austria
3. (English: 0.531) Former Chancellor To Be Charged With Perjury

   Former Chancellor Fred Sinowatz will be charged with perjury in connection with testimony in a 1987 trial that arose from a probe into President Kurt Waldheim's World War II past, the justice minister said Wednesday.
4. (English: 0.521) Austria Marks Annexation by Nazi Germany
5. (English: 0.515) Document Shows Waldheim Knew of Plan To Deport Greeks To Labor Camps
6. (English: 0.515) Documents Show Waldheim Transcribed, Forwarded Order To Kill Partisans
7. (German: 0.509) Waldheim erinnert an das "tragische Ereignis" des Jahres 1938.
8. (German: 0.507) Waldheim: Ungenaue Angaben zu Nazi-Vergangenheit gang und gäbe.
9. (English: 0.505): Austrian President Withdraws Lawsuit Against WJC President Bronfman
10. (English: 0.504) World Jewish Congress Calls Austrian Reparations To Jews "Desecration"

...

50. (French: 0.444) Publication d'un dossier sur le pass nazi de Kurt Waldheim.

We see that three German documents and seven English (AP) documents make up the top ten, all from 1988, as it turns out. The titles (or title and first paragraph) of all the stories indicate that they have to do with Waldheim. Recall that this is acheived without ever explictly matching on "Waldheim"—the term was considered separately in each of the three languages.

RUN TAG	TOPIC	LANG	TARGET
97lsiLGG	long	G	G
97lsiLFF	long	F	F
97lsiLEE	long	E	E
97lsiSGG	short	G	G
97lsiSFF	short	F	F
97lsiSEE	short	E	E

Table 3: List of monolingual runs produced

RUN TAG	TOPIC	LANG	TARGET
97lsiLGF	long	G	F
97lsiSGF	short	G	F
97lsiLFG	long	F	G
97lsiSFG	short	F	G
97lsiLEG	long	E	G
97lsiSEG	short	E	G
97lsiLEF	long	E	F
97lsiSEF	short	E	F
97lsiLGE	long	G	E
97lsiSGE	short	G	E
97lsiLFE	long	F	E
97lsiSFE	short	F	E

Table 4: List of single-language cross-language runs produced

In the returned list, we can see that the English query brought back predominantly English documents. Nonetheless, the best-matching document in the set is in German.

Catalog of Runs

For our submitted runs, we issued both long and short queries in each of the three languages against the test documents in each of three languages and returned the top 1000 for each. The actual runs are listed in Tables 3 and 4.

Preliminary results of relevance judgments are given in Table 5. The table lists, for each of the topics, the fraction of runs we submitted for which the average precision was at or above median compared to those submitted by other groups. Topics are listed in decreasing order of performance for CL-LSI.

These early results look particularly bad, especially in the monolingual case. This level of performance is far below LSI's typical performance on monolingual tasks, so we are looking for an explanation of this. It is interesting to note, however, that despite CL-LSI's overall poor early showing, it does relatively better on cross-language runs than on monolingual runs. We look forward to getting a more complete set of relevance judgments and to spending more time understanding the situations in which CL-LSI had difficulty identifying relevant documents.

Topic	Mono	Cross	Avg.
10	.33	.67	.56
6	.00	.83	.55
1	.67	.42	.50
19	.17	.58	.44
9	.00	.58	.39
14	.00	.42	.28
17	.00	.17	.11
18	.00	.17	.11
24	.00	.17	.11
5	.00	.17	.11
11	.00	.08	.05
16	.00	.08	.05
2	.00	.08	.05
average	.09	.34	.25

Table 5: Average number of runs at or above median (by topic)

References

Berry, M. W.; Dumais, S. T.; and O'Brien, G. W. 1995. Using linear algebra for intelligent information retrieval. *SIAM Review* 37(4):573–595.

Carbonell, J.; Yang, Y.; Frederking, R.; Brown, R. D.; Geng, Y.; and Lee, D. 1997. Translingual information retrieval: A comparative evaluation. In *Proceedings of Fifteenth International Joint Conference on Artificial Intelligence (IJCAI-97).*

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

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

Dumais, S. T.; Landauer, T. K.; and Littman, M. L. 1996. Automatic cross-linguistic information retrieval using latent semantic indexing. SIGIR96 Workshop On Cross-Linguistic Information Retrieval.

Dumais, S. T.; Littman, M. L.; and Landauer, T. K. 1997, to appear. Automatic cross-language information retrieval using latent semantic indexing. In Grefenstette, G., ed., *Cross Language Information Retrieval.*

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

Landauer, T. K., and Littman, M. L. 1990. 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.* 31–38.

Littman, M. L., and Keim, G. A. 1997. Cross-language text retrieval with three languages. Technical Report CS-1997-16, Department of Computer Science, Duke University.

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

Sheridan, P., and Ballerini, J. P. 1996. Experiments in multilingual information retrieval using the spider system. In Frei, H.-P.; Harman, D.; Schäble, P.; and Wilkinson, R., eds., *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96*, 58–65.