Cross-Lingual Random Indexing for Information Retrieval

Hans Moen and Erwin Marsi

Norwegian University of Science and Technology (NTNU),
Dept. of Computer and Information Science,
Trondheim, Norway
{hans.moen,emarsi}@idi.ntnu.no

Abstract. Cross-lingual information retrieval aims at retrieving relevant documents from a document collection in a language different from the query language. A novel method is proposed which avoids direct translation of queries by implicit encoding of translations in a bilingual *vector space model* (VSM). Both queries and documents are represented as vectors using an extension of *random indexing* (RI). As work on RI for information retrieval is limited, it is first evaluated for monolingual retrieval. Two variants are tested: (1) a *direct* RI model that approximates a standard VSM; (2) an *indirect* RI model intended to capture latent semantic relations among terms with a sliding window procedure. Next cross-lingual extensions of these models are presented and evaluated for cross-lingual document retrieval.

1 Introduction

In the classic *vector space model* (VSM) for *information retrieval* (IR) [26,17], both documents and queries are represented as vectors in a high-dimensional vector space. Each dimension represents term counts and terms are usually weighted using some variant of TF*IDF [10]. Relevant documents are retrieved by computing the cosine similarity between a query vector and the document vectors, retrieving the n most similar documents. A limitation of the standard VSM is that it cannot cope with semantically related terms, for example, synonyms. This was part of the motivation for *latent semantic indexing* (LSI), which uses dimensionality reduction as a means of accessing latent distributional similarities between terms [7]. Evidence for the claim that LSI improves IR seems open to interpretation. Initial evaluations suggested that LSI can improve results on certain benchmark data sets; see [3] for a summary of findings. However, more recent experimental results on a larger scale suggested otherwise [1].

Regardless of whether LSI improves retrieval or not, there is no dispute that it is computationally expensive. The core of LSI is truncated *singular value decomposition* (SVD), a mathematical operation for reducing a matrix that presumably captures higher order relations between terms. The computational cost of truncated SVD makes it hard to scale LSI to large document collections. *Random indexing* (RI), an iterative indexing method based on the principle of sparse

A.-H. Dediu et al. (Eds.): SLSP 2013, LNAI 7978, pp. 164–175, 2013.

distributed memory [11], was initially proposed as a simpler and cheaper alternative to LSI [12,23]. It is argued to deliver comparable results at a much lower computational cost. In addition, it is fully incremental, allowing addition of new documents without the need to recompute the existing model (as in LSI). It was initially evaluated for learning synonyms in a TOEFL test [12], measuring word similarity through distributional similarity, that is, through a statistical analysis of word co-occurrence frequencies in large text corpora. Since then it has been applied to a range of tasks with generally positive results [13,24]. In general, smoothing methods like LSI and RI are thought to promote a number of desirable properties in models of distributional similarity, including revealing latent meaning, reducing noise, capturing high-order co-occurrence relations, and reducing sparsity [29].

Given that LSI is claimed to improve upon the classic VSM, and that RI is claimed to be a comparable but cheap alternative to LSI, it is a logical step forward to evaluate RI in an IR context. There seem to be few studies on this. [22,5] explore RI in combination with *holographic reduced representations* (HRR). [30] use a extension of RI called *reflective random indexing* (RRI) for classifying MEDLINE articles. [2] use RI as word discrimination method in an IR task, and compare it to a word disambiguation method. [28] report results on combining RI and LSI for IR. Still, no good conclusion is given when it comes to the performance of using RI as a document index for IR. A recent review article about distributional semantics in the biomedical domain states: "To the best of our knowledge Random Indexing has not been extensively evaluated in an information retrieval context, presenting a research opportunity for its formal evaluation in the context of information retrieval from MEDLINE" [6]. The first contribution of this paper is therefore to add new empirical results on monolingual IR with RI.

Cross-lingual information retrieval (CLIR) aims at identifying relevant documents in a language other than that of the query [14]. Most approaches start with translating the query to the target language using bilingual dictionaries or machine translation systems. This raises the familiar problems in machine translation such as lack of lexical coverage and lexical translation ambiguity. Other approaches require bilingual data in the form of parallel text aligned at the word, sentence or document level. For instance, [8] propose a bilingual LSI model that requires pairs of documents and their translations for training. In contrast to existing approaches, we propose a new method called *cross-lingual random indexing* that avoids direct translation of the query. Instead translation is implicitly encoded in a RI model. There is no need for aligned bilingual text either, only a bilingual dictionary and monolingual corpora for both languages. The second contribution of this paper is therefore a new model for CLIR and its experimental evaluation.

The remainder of this paper has two major parts: random indexing for monolingual retrieval and (2) cross-lingual RI for bilingual retrieval. It concludes with a general summary of findings and an outlook on future work.

2 Monolingual Information Retrieval with Random Indexing

2.1 Direct Random Indexing

Conceptually, random indexing can be regarded as a method for compressing a standard term–document or term–term matrix, where rows (vectors) represent documents, columns represent unique terms and cells represent how many times a certain term occurs in a certain document. In practice, RI directly generates a matrix of reduced dimensionality through the following procedure:

1. Each term in the document collection gets a unique *index vector*. Index vectors are high-dimensional, but typically of substantially lower size than the total number of unique terms. These are very sparse randomly initiated vectors containing mostly zeros, apart from a few randomly chosen 1s and -1s. As a result these index vectors becomes "nearly orthogonal" to each other in the vector space.
2. Each document is then represented by a *document vector* obtained by summing the index vectors of all terms occurring in the document. This optionally includes term weighting and vector length normalization.

As a result of this procedure, documents containing the same terms have vectors composed of the same index vectors and are therefore more similar in the vector space. The vector for a query likewise is constructed by summing the (weighted) index vectors of all its terms.

2.2 Indirect Random Indexing

Indexing a text corpus with *sliding window* RI takes a somewhat different approach [13]. Instead of directly summing the index vectors of a document's terms, there is an intermediate step that first creates *term context vectors*. Indirect RI involves the following steps:

1. Each unique term in the document collection gets a unique *index vector*.
2. Next a *context vector* is generated for each term. The document collection is scanned by sliding a fixed-size window over the text, term by term. Each step, the context vector of the term in the center of the window – often referred to as the *target term* – is updated by adding the index vectors of the neighboring terms within the window. As a result, terms co-occurring with similar terms obtain similar context vectors in the vector space.
3. Context vectors are normalized by dividing them by the global frequency of the term in the document collection.
4. Each document is then represented by a *document vector* obtained by summing the context vectors of all its terms, optionally including term weighting and vector length normalization.

This method thus models higher-order co-occurrence relations among terms, captured through analyzing local co-occurrence relations among words. In addition, there are methods for encoding word order relations within the sliding window. These options and other experimental variables are detailed in the next section.

2.3 Experimental Setup

We adopted the well-established CLEF framework for evaluation of cross-lingual information retrieval, in particular, the ad hoc monolingual and bilingual tracks from CLEF 2005 [4]. Monolingual experiments address English, whereas bilingual experiments concern German and English as source and target language respectively. This choice was primarily prompted by our access to CLEF 2004-2008 data, as well as to a relatively large German-English translation dictionary. The CLEF data consists of three components: document collections, search topics, and relevance judgement; see [4] for details.

Document collections comprise news text from news wires, newspapers and opinion magazines. Stopword removal and lemmatization were applied as this was found to generally improve the IR scores. The full English corpus consist of 257,130 documents with approximately 130M words, and among these 325,617 unique ones after lemmatization. After stopword removal, the corpus is reduced to 70M words, with 325,392 unique words. All documents were used for training, and a subset of 169,477 was used as retrievable documents in the experiment. Documents were lemmatized with TreeTagger. Stopwords were removed using customized versions of the default stopword lists provided by the Lucene project [16]. Terms occurring only once were also removed. For all remaining terms, TF*IDF values were calculated [10], and used for weighting terms, i.e. their context vectors, when creation of document vectors.

Topics express the informational need of a user and consist of three fields: (1) a brief *title* stating the main keywords, (2) a single sentence *description* of the concept conveyed by the keywords, and (3) a more elaborate narrative. All experiments in this paper used the combination of *title* and *description* to create a query.

Relevance judgements specify which documents from the document collection are relevant to a particular topic. Documents are assessed as either relevant or irrelevant to the topic by a panel of human judges.

The RI algorithms used for the experiments in this paper are based on the JavaSDM package [9]. Scores are calculated using the `trec_eval` tool (version 7.3). Results are reported in terms of mean average precision (MAP) together with the total number and percentage of relevant documents retrieved over all 50 queries. For comparison we used Apache Lucene [16] (v4.1.0), a state-of-the-art search engine implementing a TF*IDF weigthed variant of the standard VSM. No additional weighting or "boosting" of specific sections or fields in the documents or queries were applied.

Experiments explored a number of different configurations. The first two parameters concern the RI model itself:

Dimensionality. The size of the vectors (index and context vectors) ranged from 1000 through 1800 to 4000.

Non-zeros. The total number of 1's and -1's randomly assigned to the index vectors.

In addition, there were two parameters that only apply to Indirect RI:

Window Size. The size of the sliding window ranged from 2+2 (i.e. two words on the left and two words on the right of the target term) up to 20+20.

Weighting Scheme. Index vectors of the neighboring terms in the sliding window are weighted and/or modified before they are added to a context vector. *distance weighting* uses the function $2^{1-distance}$, *distance* being the distance in words to the target term [13]. *Random permutations* (RP) [25] encode word order relations by shifting the elements in the index vectors according to both their position and their distance from the target term. In a similar fashion, *Direction vectors* only encode direction by shifting index vectors once, either left or right depending on which side of the target term they are located [25], plus weighting the vectors similarly as in Distance Weighting.

2.4 Results

Lucene retrieved 1817 relevant documents (88.08%), resulting in a a MAP score of 0.3713. Table 1 presents corresponding results for direct RI, indicating that about 64–72% of the relevant documents were found, with a MAP score in the range from 0.15–0.18. Increasing the number of non-zeros up till 8 was found to improve results while changing dimensionality had no effect.

Table 2 presents selected results for Indirect RI. Vector dimensionality does not affect the results beyond a certain size, approximately around 2000. The number of non-zeros also has little effect, less so than in the Direct RI experiments. Larger window sizes appear to yield better results than smaller sizes. Weighting schemes do not have any positive effect, suggesting that word order within the window is irrelevant. Smaller window sizes were tested for the other weighting schemes, but none of these performed better than without weighting. In sum, a medium vector dimensionality (1800) together with a large window size (16+16), unweighted, and few non-zeros (4) gave the best performance.

2.5 Discussion

The direct RI method is essentially an approximation of the standard TF*IDF-weighted VSM. However, where the VSM would have a dimensionality equal to the number of unique terms in the document collection (e.g. 325,617 for English), direct RI has just 1800, which amounts to approximately 2% of the size. This may explain why Direct RI scores lower than what may be expected from a standard VSM, here represented by Lucene.

Table 1. Results with direct random indexing for monolingual (English) ad hoc information retrieval track from CLEF 2005

Dimensions	Non-zeros	MAP	Found/2063	%Found
1800	2	0.1512	1340	64.95
1800	4	0.1769	1427	69.17
1800	**8**	**0.1839**	**1481**	**71.79**

Table 2. Results with indirect random indexing for monolingual (English) ad hoc information retrieval track from CLEF 2005

Dim.	Non-zeros	Window	Weighting	MAP	Found/2063	%Found
1800	4	2+2	No weighting	0.1411	1238	60.01
1800	4	4+4	No weighting	0.1722	1316	63.79
1800	4	8+8	No weighting	0.1920	1387	67.23
1800	4	12+12	No weighting	0.1965	1415	68.59
1800	4	**16+16**	No weighting	0.1987	**1426**	**69.12**
1800	4	20+20	No weighting	0.1984	1420	68.83
1800	2	16+16	No weighting	0.1954	1413	68.49
1800	**4**	16+16	No weighting	0.1987	**1426**	**69.12**
1800	8	16+16	No weighting	0.1965	1400	67.86
1000	4	16+16	No weighting	0.1961	1400	67.86
1800	4	16+16	No weighting	0.1987	**1426**	**69.12**
4000	4	16+16	No weighting	**0.1998**	1411	68.40
1800	4	16+16	Rand. Permutations	0.1422	1067	51.72
1800	4	16+16	Direction Vectors	0.1391	1221	59.19
1800	4	16+16	Dist. weighting	0.1477	1286	62.34
1800	4	16+16	**No weighting**	0.1987	**1426**	**69.12**

We also find that indirect RI achieves slightly better mean average precision than direct RI, suggesting a better ranking among the top 1000 retrieved documents, whereas direct RI yields better recall. This finding is in agreement with the conclusions in [22]. Differences are small though (2.67%) and this may therefore cast some doubt on the claim that the sliding window variant captures latent relations between terms. Alternatively, it may be interpreted as an indication that modeling latent semantic information does not consistently improve the IR results. In fact, some recent studies suggest that LSI also yields poor retrieval accuracy on a large number of TREC bench mark sets [1].

3 Cross-Lingual Information Retrieval with Random Indexing

3.1 Method

The core idea in the method for cross-lingual RI proposed here is that source and target language models share the same vector space. In this way, the vector representation of a query stated in the source language can be compared directly to the vector representation of documents in the target language. This removes the need for any explicit translation, as term translations and cross-lingual synonymy are implicitly encoded in the vector space. This is accomplished through a sharing of index vectors across languages during the random indexing procedure, so that

terms that are translations of each other share a common index vector. Two variants of direct and indirect cross-lingual RI based on this idea are detailed below.

As a baseline for comparison, we use the dictionary to translate the queries, translating each source term into the corresponding $top-N$ most frequent target terms according to the TL corpus. In addition, terms not in the dictionary are simply copied over, assuming a lot of these are proper nouns. These translated queries are then used by Lucene for monolingual IR in the TL.

Direct Cross-Lingual Random Indexing. The method for cross-lingual direct RI is almost the same as for monolingual direct RI (cf. Section 2.1), except for one crucial modification in the first step, where index vectors are shared across languages. This assumes a translation dictionary mapping source language terms to target language terms, with one-to-many mappings in the case of translation ambiguity. First, a unique index vector is generated for each source term in the dictionary. Next, each target term gets the same index vector as its corresponding source term. If a target term serves as the translation of multiple source terms, their index vectors are merged with disjunction. The second step of creating query and document vectors is the same as for monolingual RI.

Indirect Cross-Lingual Random Indexing. As in the direct cross-lingual case, index vectors are again shared among source terms and their translations. Source language and target language document collections are then processed independently using the sliding window procedure to build term context vectors for source and target language terms respectively (step 2), followed by frequency correction (step 3). Notice that documents are not aligned in any way and are in fact completely unrelated. Finally, (multilingual) document vectors are obtained by summing the context vectors of all target terms contained in the document, whereas query vectors are constructed by summing vectors for their source terms.

A variant of this approach includes an extra step following the construction of the term context vectors. For each context vector of a source term, we add to it all the context vectors of its translations. Conversely, for each context vector of a target term, we add to it all the context vectors of the source terms it is a translation of. The resulting enriched context vectors will be referred to as *translation vectors*. The reasoning behind this operation is that translation vectors presumably encode second-order translation relations. That is, a pair of vectors representing source and target language texts is not only similar when the texts contain terms that are translations of each other, but also when the texts contain terms co-occurring with terms that are in turn translations of each other. This is akin to query expansion through related terms used to improve recall.

3.2 Experimental Setup

A proprietary German-English translation dictionary was used in the process of constructing index vectors. It is lemma-based, provides part-of-speech (POS)

tags on both source and target side, and contains over 576k entries. In experiments we only used single-word expressions, leaving out the multi-word expressions, which did not seem to be benificial.

Cross-lingual experiments were based on the bilingual ad hoc retrieval track using 50 German topics to retrieve English documents. The German topics corresponded to the English topics used earlier in the monolingual experiments; the English document collection was the same as before (cf. Section 2.3). However, the use of a translation dictionary imposed some additional constraints. First, the dictionary entries are lemma-based, so for the purpose of look-up, the document collections were lemmatized with TreeTagger using pre-trained models for English and German [27]. Two variations where tested, one including out-of-dictionary terms during training, and one where terms were limited to those in the dictionary. For the latter, this reduced the number of unique English terms from 325,617 to 114,645 and the total number of indexed terms in the English document collection from approximately 130M down to 70M (after stopword removal). Likewise, the number of unique German terms was reduced from 1,057,526 to 144,766 and the total number of indexed terms from about 80M down to 36M.

Model parameters were adopted from the best scoring configurations in the monolingual IR experiment presented earlier: a vector dimensionality of 1800, 4 non-zeros for index vectors, and a unweighted window of 16+16 in indirect RI.

3.3 Results

Table 3 shows results for applying the cross-lingual random indexing method to the *bilingual* ad hoc IR track. These scores are clearly a lot lower than the monolingual scores, with direct RI again outperforming indirect RI in terms of MAP scores. However, the variant of indirect RI employing *translation vectors* performs best in terms of recall. The latter was also tested using out-of-dictionary terms, resulting in lower recall but higher MAP. Unfortunately none of the RI methods were able to beat the baseline relying on a two-step approach of query translation using the dictionary followed by monolingual IR with Lucene. As shown in Table 4, best MAP and recall scores were obtained by taking the two or three most frequent translations respectively.

Table 3. Results with cross-lingual random indexing for bilingual (German-English) ad hoc information retrieval track from CLEF 2005

Method	MAP	%Mono	Found/2063	%Found
Direct Cross-lingual RI	**0.0667**	**36.27**	592	28.70
Indirect Cross-lingual RI	0.0176	8.56	400	19.39
Translation vectors limited to dictionary	0.0501	25.21	**767**	**37.18**
Translation vectors not limited to dictionary	0.0656	33.02	659	31.94

Table 4. Results for combination of query translation and Lucene on bilingual (German-English) ad hoc information retrieval track from CLEF 2005

Method for query translation	MAP	%Mono	Found/2063	%Found
Dictionary ranked top1	0.1436	38.68	978	47.41
Dictionary ranked top2	**0.1541**	41.50	1068	51.77
Dictionary ranked top3	0.1437	38.70	**1091**	**52.88**
Dictionary ranked top4	0.1347	36.28	1073	52.01
Dictionary ranked top5	0.1275	34.34	1036	50.22

3.4 Discussion

Among the participants in CLEF 2005, none of them submitted any results for English-German in the ad hoc bilingual track. However, three teams targeted English from other source languages. University of Glasgow submitted results for Greek-English [15]. After query expansion, Greek lemmas were automatically translated into English with Yahoo's Babelfish, a full fledged MT system. The best results were obtained with the classic BM25 model [21] with empirically tuned parameter. They achieved a MAP score of 0.2935, 68.14% of their reported monolingual score. Johns Hopkins University worked on Greek-English, Hungarian-English and Indonesian-English, aiming for a language-independent solution based on character n-grams [18]. Queries were expanded prior to translation using the source language CLEF corpus. Next, queries were translated using online translation services: Yahoo's Babelfish for Greek, ToggleText's Kataku for Indonesian and TranslationExpert's InterTran for Hungarian. A statistical language model was employed for retrieval. They achieved MAP scores of 0.2418 (54.94%) for Greek, 0.3728 (84.71%) for Indonesian, and 0.1944 (44.17%) for Hungarian. University of Indonesia reported results for Indonesian-English. Queries were first translated using Transtool, a commercial MT system. Retrieval relied on VSM using the Lucene IR system, with a best MAP score of 0.1830 (52.16%).

Scores obtained with the cross-lingual RI methods are thus relatively low compared with other approaches using generic MT systems for translating the query prior to monolingual retrieval. We believe that the same issues that make the RI model score quite a bit lower than the full VSM in monolingual IR, are also present in the cross-lingual RI method tested here, together with other factors such as dictionary coverage.

There is some related work on the notion of bilingual vector spaces. Most related is the work by Dumais et al [8], who proposed a model for cross-lingual IR based on bilingual LSI. In contrast to the cross-lingual RI methods, their approach requires an aligned corpus of documents and their translations for training purposes. In a different area, Rapp proposed cross-lingual distributional similarity formalized as bilingual vector spaces to identify translation pairs in non-parallel text [20]. Peirsman & Padó used a bilingual vector space as an intermediary step in a model for learning selectional preferences [19]. Sahlgren & Karlgren describe an approach for automatic extraction of bilingual lexica using random indexing of parallel corpora [24].

4 Conclusion and Future Work

The first contribution of this paper is experimental results for random indexing in document retrieval by applying it to the monolingual (English) ad hoc IR track from CLEF 2005. It was found that indirect RI, which uses a sliding window approach during training, achieves slightly better mean average precision than direct RI, which is conceptually a compressed version of a standard VSM, suggesting a better ranking among the retrieved documents, whereas direct RI yields slightly better recall. A full VSM model as implemented in Lucene achieved better results than both of these. This is inconsistent with the claim that models such as LSI and RI improve retrieval because they model latent semantic relations among terms.

The second contribution is a new method for cross-lingual RI in which source and target language models share the same vector space, allowing direct comparison of the vector representations of source and target language texts without the need for any explicit translation. This is accomplished through a sharing of index vectors across languages during the random indexing procedure. It requires a translation dictionary and unrelated monolingual text corpora, but no aligned bilingual text. Of the three different variants proposed, indirect cross-lingual RI with translation vectors performed best when applied to the German-English bilingual ad hoc IR track from CLEF 2005. A straight-forward method of using a dictionary for translation of the queries and then Lucene for monolingual IR achieved better results than using our proposed methods.

Despite relatively low performance, the cross-lingual RI approach may still be attractive because of several advantages. First, it is very light-weight in terms of resources, as it only requires a translation dictionary. There is no need for bilingual data in the form of parallel documents or word-aligned text, which can be expensive to construct. Second, it inherits the computational simplicity from standard RI and is therefore scalable to huge document collections while retaining relateively small models. Third, additional target languages can be added without the need to retrain the existing models. Forth, queries and documents are in the same cross-lingual vector space, so no explicit translation step is required. In addition, the method may have potential uses in specialized domains utilizing specialized sublanguages where little or no aligned training data is available. One such example being the clinical domain, which contains specialized documents for which parallel or aligned text is difficult to produce and obtain. This may also include *cross-domain IR*, possibly incorporating domain knowledge into the cross-language/-domain dictionary to model domain-dependent relations among terms and documents.

There are still many unsolved questions related to application of RI in retrieval. For instance, no good explanation is yet given for why capturing latent semantic relations among terms seemingly does not improve document retrieval. One possible explanation is that the features which make two documents similar, or dissimilar, are not the same as those that determine similarity on a term level (e.g. synonymy). Another explanation could be that the way vectors are combined into documents, i.e. through TF*IDF weighted summation, is not optimal

for capturing or preserving higher-order semantic relations among terms. More experimentation is needed to explore a wider range of model configurations on more benchmark data, both for monolingual IR using vectors of higher order semantic information and for CLIR with language pairs other than German–English. A direct comparison between LSI and RI for CLIR is desirable as well. There is also a need for a more thorough evaluation of using the presented *term translation vectors* in detecting semantically similar terms across languages.

Acknowledgements. We would like to thank Björn Gambäck for discussion and useful suggestions. This work was partly funded by the EviCare project (http://www.evicare.no) and by the European Community's Seventh Framework Programme (FP7/20072013) under grant agreement nr 248307 (PRESEMT).

References

1. Atreya, A., Elkan, C.: Latent semantic indexing (LSI) fails for TREC collections. SIGKDD Explorations 12(2), 5–10 (2010)
2. Basile, P., Caputo, A., Semeraro, G.: Semantic vectors: an information retrieval scenario. In: IIR, pp. 27–28 (2010)
3. Berry, M., Dumais, S., O'Brien, G.: Using linear algebra for intelligent information retrieval. SIAM Review 37(4), 573–595 (1995)
4. Braschler, M., Peters, C.: CLEF Methodology and Metrics. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2001. LNCS, vol. 2406, pp. 394–404. Springer, Heidelberg (2002)
5. Carrillo, M., Villatoro-Tello, E., López-López, A., Eliasmith, C., Montes-y-Gómez, M., Villaseñor-Pineda, L.: Representing context information for document retrieval. In: Andreasen, T., Yager, R.R., Bulskov, H., Christiansen, H., Larsen, H.L. (eds.) FQAS 2009. LNCS, vol. 5822, pp. 239–250. Springer, Heidelberg (2009)
6. Cohen, T., Widdows, D.: Empirical distributional semantics: Methods and biomedical applications. Journal of Biomedical Informatics 42(2), 390 (2009)
7. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science 41(6), 391–407 (1990)
8. Dumais, S., Letsche, T., Littman, M., Landauer, T.: Automatic cross-language retrieval using latent semantic indexing. In: AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, pp. 15–21 (1997)
9. Hassel, M.: JavaSDM package (2004), `http://www.nada.kth.se/~xmartin/java/`
10. Jones, K.S.: A Statistical Interpretation of Term Specificity and its Application in Retrieval. Journal of Documentation 28(1), 11–21 (1972)
11. Kanerva, P.: Sparse distributed memory: A study of psychologically driven storage. MIT press (1988)
12. Kanerva, P., Kristoferson, J., Holst, A.: Random indexing of text samples for latent semantic analysis. In: Gleitman, L., Josh, A. (eds.) Proceedings of the 22nd Annual Conference of the Cognitive Science Society, p. 1036. Erlbaum, Mahwah (2000)
13. Karlgren, J., Sahlgren, M.: From Words to Understanding. In: Uesaka, Y., Kanerva, P., Asoh, H. (eds.) Foundations of Real-World Intelligence, pp. 294–308. CSLI Publications, Stanford (2001)

14. Kishida, K.: Technical issues of cross-language information retrieval: a review. Information Processing & Management 41(3), 433–455 (2005)
15. Lioma, C., Macdonald, C., He, B., Plachouras, V., Ounis, I.: Applying Light Natural Language Processing to Ad-Hoc Cross Language Information Retrieval. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 170–178. Springer, Heidelberg (2006)
16. Apache Lucene open source package, `http://lucene.apache.org/`
17. Manning, C., Raghavan, P., Schütze, H.: Introduction to information retrieval, vol. 1. Cambridge University Press, Cambridge (2008)
18. McNamee, P.: Exploring New Languages with HAIRCUT at CLEF 2005. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 155–164. Springer, Heidelberg (2006)
19. Peirsman, Y., Padó, S.: Cross-lingual Induction of Selectional Preferences with Bilingual Vector Spaces. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 921–929. Association for Computational Linguistics, Los Angeles, Los Angeles (2010)
20. Rapp, R.: Identifying word translations in non-parallel texts. In: Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics, pp. 320–322. Association for Computational Linguistics (1995)
21. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford., M.: Okapi at trec-3. In: Proceedings of the Third Text REtrieval Conference (TREC 1994), Gaithersburg, USA (1994)
22. Ruiz, M., Eliasmith, C., López, A.: Exploring the Use of Random Indexing for Retrieving Information. Tech. Rep. CCC-08-006, INAOE (2008)
23. Sahlgren, M.: An introduction to random indexing. In: Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE, vol. 5 (2005)
24. Sahlgren, M., Karlgren, J.: Automatic bilingual lexicon acquisition using random indexing of parallel corpora. Natural Language Engineering 11(03), 327–341 (2005)
25. Sahlgren, M., Holst, A., Kanerva, P.: Permutations as a Means to Encode Order in Word Space. In: Proceedings of the 30th Conference of the Cognitive Science Society (2008)
26. Salton, G., Wong, A., Yang, C.: A vector space model for automatic indexing. Communications of the ACM 18(11), 613–620 (1975)
27. Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, vol. 12, pp. 44–49 (1994)
28. Sellberg, L., Jönsson, A.: Using random indexing to improve singular value decomposition for latent semantic analysis. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008) (May 2008)
29. Turney, P., Pantel, P.: From Frequency to Meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research 37, 141–188 (2010)
30. Vasuki, V., Cohen, T.: Reflective random indexing for semi-automatic indexing of the biomedical literature. Journal of Biomedical Informatics 43(5), 694–700 (2010)