# Towards Cross-Lingual Information Retrieval using Random Indexing

Hans Moen, Erwin Marsi
Norwegian University of Science and Technology,
Department of Computer and Information Science, 7491 Trondheim, Norway
hans.moen@idi.ntnu.no, emarsi@idi.ntnu.no

**Abstract**

We propose a new method for cross-lingual information retrieval that avoids direct translation of queries by implicit encoding of translations in a Random Indexing vector space model. It relies solely on a translation dictionary and source/target language text corpora, without the need for aligned multilingual text. Initial experiments show promising results for retrieval of related terms in Norwegian-English cross-lingual information retrieval. A number of possible improvements and extensions to the model are discussed.

## Introduction

Cross-lingual information retrieval (CLIR) aims at identifying relevant documents in a language other than that of the query (Kishida, 2005). Most approaches focus on translating the query to the targeted 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. In this paper we present preliminary work on an approach to CLIR that avoids direct translation of the query. Instead translation is implicitly encoded in the vector space model used for information retrieval (IR). For this purpose we use Random Indexing (RI) (Kanerva et al., 2000), an iterative indexing method based on the principle of sparse distributed memory (Kanerva, 1988). RI exhibits several desirable properties that lack in more conventional vector/word space models (VSMs) (Sahlgren, 2005). RI incrementally builds a vector space representing either words or documents, where the cosine between vectors serves a measure of their similarity. Similarity relations between the elements in the vector space, words or documents, represent contextual similarity, based on statistical word co-occurrence information.

Indexing a text corpus with sliding window RI comprises the following steps:

1. Each term in the corpus is given a randomly initialized unique *index vector*. These index vectors are sparse and high-dimensional, containing a small amount of 1's and -1's.
2. A *context vector* for each term is generated. This is done by going through the text term-by-term with a window of fixed size (usually 4 or 6 terms wide) over the focused term. The index vectors of the other terms in the window are *added* to the context vector of the focused term. As a result, terms having co-occurred with similar neighboring terms obtain similar context vectors in this vector space.
3. In a similar vein, context vectors for sentences or documents can be constructed by summing the context vectors of all terms contained in them (possibly weighted using some term weighting scheme).

Retrieval is then accomplished as follows:

4. A context vector for the query is constructed by adding the context vectors of all its terms.
5. Documents in the corpus are ranked according to the cosine similarity between the query's context vector and the context vectors for each document.

*This paper was presented at the NIK 200x conference. For more information see http://www.nik.no/*

The CLIR extension proposed here is motivated by the following hypothesis: "If – for example – in a text of one language two words A and B co-occur more often than expected by chance, then in a text of another language those words that are translations of A and B should also co-occur more frequently than expected." (Rapp, 1995). The CLIR method includes the same operations as above for indexing a corpus, but with one crucial difference. As in step 1 above, each source language term gets a unique index vector. Then, however, a bilingual dictionary is used to lookup its translation and the corresponding target language term get the *same* index vector as the source term. These shared index vectors form the basis for measuring cross-lingual similarity between documents in different languages. We will address the possibility of multiple possible translations further on. Steps 2 and 3 are then carried out as before per language using monolingual corpora.

The CLIR step is performed much like how it is done for monolingual RI, but instead of comparing context vectors within the same language, we now directly compare context vectors in one language with context vectors in the other language using the cosine similarity function. It is worth noting that this is done just as easily as for monolingual RI; no additional steps are needed after the training. Also, new languages can be added without the need to retrain the existing models. Yet another advantage is that no word-aligned bilingual corpora are needed.

## Related work

There is a substantial body of related work on creating or extending translation dictionaries. Automatic extraction of bilingual dictionaries is typically performed on word-aligned bilingual corpora. However, these have a limited availability and are expensive to build from scratch. Rapp (1995), Kikui (1999), Fung and Yee (1998), and Chiao and Zweigenbaum (2002) proposed methods for automatically learning new word translations from comparable but non-parallel monolingual text, based on the assumption that a term and its translation appear in similar contexts. The context of terms is also modeled as vectors in a vector space model, but separate models are built for source and target language. Computing similarity between a source vector and target vectors therefore involves a preliminary step of translating the words/features in the context vector. Others have applied similar methods but with a focus on CLIR (Picchi and Peters, 1998). The idea of a bilingual vector space is also introduced in (Peirsman and Padó, 2010), although only for bootstrapping a dictionary from cognates as a sub-task in learning selectional preferences.

Latent Semantic Analysis (LSA) (Deerwester et al., 1990) is a method based on VSM representation. It has been applied in various CLIR and translation methods termed Cross-Language Latent Semantic Indexing (CL-LSI) (Dumais et al., 1997). These methods use aligned cross-language documents, either aligned on term level or on document level. RI has also been used for translation (Sahlgren et al., 2003; Karlgren et al. 2005). RI has, as mentioned earlier, several desirable properties over LSA, and the most important ones are that it is incremental and scales better to large corpora. The method described by Karlgren et al. (2005) uses aligned documents. A common index vector is assigned to each aligned document pair. Context vectors for all terms in each language are then trained/generated by adding the index vectors from the document pairs where they occur. Despite good results, these methods require aligned cross-lingual documents for training, either term-aligned or document-aligned.

## Initial Evaluation

To get an idea of how the method performs on a CLIR task, we conducted a pilot-experiment attempting to retrieve semantically related terms across two languages. For this experiment we used a Norwegian-English dictionary containing approximately 45400 term pairs. We limited the possible terms to only those found in the dictionary. As text material, we randomly selected one million Norwegian and one million English sentences from large corpora of web text for the respective languages. All terms in the corpora and the lexicon where lemmatized. For creating context vectors, a window size of four terms was employed (two anterior and two posterior terms, weighted by their distance from the target term). The dimensionality of index and context vectors was 1024. The number of non-zeros in the random index vectors was four. The JavaSDM package (http://www.nada.kth.se/~xmartin/java) served as the basis for our implementation.

One complication with creating index vectors for translation pairs is that a term can be ambiguous. Not only can a term have multiple meanings in the source language, a term in one language can also map to multiple terms in another language (according to the dictionary). As a first attempt, we opted for a "source language centered" mapping of all translations found in the dictionary. First, an index vector was created for each unique Norwegian term in the dictionary. Next, for every English translation of a Norwegian term, the Norwegian term's index vector is added to the index vector of the corresponding English terms. In doing so we ensure that all translation relations are preserved in the vector space models. However, terms in the target language that map to multiple terms in the source language will get multiple sets of non-zeros in their index vectors, which in turn, according to the cosine similarity function, decreases their similarity to each of their counterparts in the source language. There are alternative ways of performing this mapping process, as will be discussed below.

The training of the model was done as described earlier. We then used Norwegian terms as queries and retrieved the top 10 English terms selected as most similar by the model. Some results are shown in Table 1.

| *Query:* | *Translation (dictionary):* | *Retrieved:* |
|---|---|---|
| gå | walk, go, pass, go by, leave, be on, be performed, happen, turn out | eat, go, suck, come, perish, catch fire, exist, resign, flourish, disintegrate |
| far | father, track, trail | aunt, uncle, grandmother, wife, waistcoat, sister, co-worker, sister-in-law, dad, daughter-in-law |
| skomaker | shoemaker, bootmaker, cobbler | philanthropist, internist, trapper, grow old, aunt, yellow fever, fortune teller, tradesman, witty, vagabond |

Table 1 Term retrieval results for Norwegian-English CLIR on similar terms.

## Discussion and Future Work

The initial retrieval results from our model seem to be generally sensible but noisy. This does not necessarily diminish its value with respect to CLIR. So far we have not tried our approach for sentence or document retrieval. However, context vectors for sentences and documents can be created simply by summing the context vectors for all terms they contain. It also makes sense to weight terms (e.g. using TF*IDF). This should make future experiments on CLIR benchmark data sets fairly easy.

It should also be noted that the source and target text corpora used so far are completely unrelated random samples from a web corpora. The use of comparable corpora from related domains instead may improve retrieval results.

The issue of translation ambiguity may be reduced by exploiting part-of-speech information in the cross-language term-matching process, meaning we can have separate index vectors for most translation options, which in turn may help to tease apart contexts that are currently conflated.

Another interesting direction for future work is the use of different *domain-specific* corpora. For two such corpora in the same language, one could apply a *cross-domain dictionary* instead of a cross-language dictionary to enable *cross-domain IR*.

Norwegian & English have fairly similar grammar, but this is not the case for many other language pairs. Further work should aim to find optimal parameters related to the sliding window and how context vectors are created, i.e. its size and weighting scheme.

In the initial experimentation conducted for this paper we did not allow the system to generate context vectors for terms other than those found in the dictionary. However, allowing these out-of-dictionary terms to get context vectors generated for them could possible also work, enabling these to effect the retrieval result.

### Acknowledgments

This work was partly funded by the EviCare project (http://www.evicare.no) and by the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement nr 248307 (PRESEMT). We would also like to thank the reviewers for their valuable comments.

## References

Chiao, Y., & Zweigenbaum, P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. *Proceedings of the 19th international conference on Computational linguistics*, *2*, 3-7.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American society for information science*, *41*(6), 391–407. Citeseer.

Dumais, S., Letsche, T., & Littman, M. (1997). Automatic cross-language retrieval using latent semantic indexing. *Proceedings of the AAAI Symposium on Cross-Language Text and Speech Retrieval* (pp. 18-24).

Fung, P., & Yee, L. Y. (1998). An IR Approach for Translating New Words from Nonparallel, Comparable Texts. *Proceedings of the 17th international conference on Computational linguistics-Volume 1* (pp. 414–420). Association for Computational Linguistics.

Kanerva, P. (1988). Sparse distributed memory. *The MIT Press*.

Kanerva, P., Kristoferson, J., & Holst, A. (2000). Random indexing of text samples for latent semantic analysis. *In Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pp.1036.

Karlgren, J., Sahlgren, M., Järvinen, T., & Cöoster, R. (2005). Dynamic lexica for query translation. *Lecture Notes in Computer Science*, *3491*(Multilingual Information Access for Text, Speech and Images), 150-155.

Kikui, G. (1999). Resolving translation ambiguity using non-parallel bilingual corpora. *Proceedings of ACL99 Workshop on Unsupervised Learning in Natural Language Processing* (pp. 31-36).

Picchi, E., & Peters, C. (1998). Cross-language information retrieval: A system for comparable corpus querying. *Grefenstette (Grefenstette, 1998a), chapter 7,* (pp. 81–90).

Peirsman, Y., & Padó, S. (2010). Cross-lingual induction of selectional preferences with bilingual vector spaces. *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.

Rapp, R. (1995). Identifying word translations in non-parallel texts. *Proceedings of the 33rd annual meeting on Association for Computational Linguistics* (pp. 320–322). Association for Computational Linguistics.

Sahlgren, M., Karlgren, J., Cöoster, R., & Jäarvinen, T. (2003). SICS at CLEF 2002: Automatic query expansion using random indexing. *Lecture Notes in Computer Science*, *2785*(Advances in Cross-Language Information Retrieval), 311-320.

Sahlgren, M. (2005). An introduction to random indexing. *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005* (pp. 1-9). Citeseer.