Lightweight Random Indexing for Polylingual Text Classification (Extended Abstract)*

Alejandro Moreo Fernández, Andrea Esuli and **Fabrizio Sebastiani**
Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche
{alejandro.moreo,andrea.esuli,fabrizio.sebastiani}@isti.cnr.it

Abstract

Polylingual Text Classification (PLC) is a supervised learning task that consists of assigning class labels to documents written in different languages, assuming that a representative set of training documents is available for each language. This scenario is more and more frequent, given the large quantity of multilingual platforms and communities emerging on the Internet. In this work we analyse some important methods proposed in the literature that are machine-translation-free and dictionary-free, and we propose a particular configuration of the Random Indexing method (that we dub *Lightweight Random Indexing*). We show that it outperforms all compared algorithms and also displays a significantly reduced computational cost.

1 Introduction

With the rapid increase in the amount of multi-cultural and multilingual information on the Internet, how to properly classify texts potentially belonging to many different languages has become a problem of relevant practical interest. How to effectively leverage multilingual information constitutes a sub-task of *Multilingual Text classification* (MLC) typically known as *Polylingual Text Classification* (PLC). In PLC, and differently from *Cross-Lingual Text Classification* (CLTC) [Bel *et al.*, 2003], a representative set of training documents for each language is assumed available. Therefore, a straightforward solution may consist of training a separate monolingual classifier for each language (the so-called *naïve polylingual classifier* [García Adeva *et al.*, 2005]) and merging the outcomes of each classifier [Amini *et al.*, 2009]. However, such a solution is suboptimal, since each classifier does not take advantage of the training documents available for the other languages. The challenge in PLC consists thus of improving the final classification performance with respect to a set of independent monolingual classifiers.

In addition to traditional difficulties encountered in TC, there are specific obstacles that arise in the polylingual scenario, which are mainly related to the *high dimensionality* problem (which is due to the presence of multiple languages) and the *feature disjointness* problem (which is due to their lack of lexical overlap).

To overcome these difficulties some authors propose the use of automatic machine translation (MT) tools [Bel *et al.*, 2003; Wei *et al.*, 2011], multilingual ontologies [Nastase and Strapparava, 2013], or bilingual corpora [Vinokourov *et al.*, 2002] as a means to fill the gaps among the different languages. These approaches are however limited by the need of external resources, which might not be available for all languages of interest or not be public / free to use. As a response to these drawbacks, methods for the automatic acquisition of bilingual dictionaries [Wei *et al.*, 2014] have been proposed. However, these methods are affected by high computational costs, due to the use of sophisticated statistical analysis. With the goal of reducing the restrictions imposed to the final classifier, we will restrict our investigation to methods that are both MT-free and dictionary-free.

One of the main challenges in PLC concerns the relevant increase in the number of features that represent the documents. An important dimensionality reduction technique is Latent Semantic Analysis (LSA) [Deerwester *et al.*, 1990], which has been later applied to cross-lingual problems [Dumais *et al.*, 1997] and multilingual classification [Xiao and Guo, 2013]. *Random Indexing* (RI) [Kanerva *et al.*, 2000; Sahlgren, 2005], a method belonging to the family of Random Projections methods [Papadimitriou *et al.*, 1998], arises as a computationally cheaper alternative to LSA [Sahlgren, 2001], while at the same time preserving some important characteristics of LSA [Fradkin and Madigan, 2003].

In this work we investigate the suitability of RI as a representation scheme for PLC. Besides the fact that the original relative distances are approximately preserved in the new space [Johnson *et al.*, 1986], it turns out that, in a multilingual scenario, each latent axis is defined as a random linear combination of the original terms regardless of the language these terms belong to. As a result the entire new space becomes informative for all languages at once. While RI has already been tested in multilingual scenarios [Sahlgren and Karlgren, 2005] and monolingual TC [Sahlgren and Cöster, 2004], to the best of our knowledge it has not been tested in the PLC case so far. We identify a particular configuration of RI, dubbed *Lightweight Random Indexing* (LRI) [Moreo Fernández *et al.*, 2016], that presents a significantly

*This paper is an extended abstract of an article appeared as [Moreo Fernández *et al.*, 2016].

reduced computational cost, and we conduct an analytical study that can be useful to better understand the nature of random mapping methods.

2 Lightweight Random Indexing

RI is a distributional semantic model that builds *distributional vectors* for words based on the observation of the terms they co-occur with. To do so, RI maintains a dictionary of *random index* vectors for each feature in the original space. Each random index vector consists of an n-dimensional sparse vector with k non-zero values, randomly distributed across +1 and -1. The distributional vector of a word is iteratively updated by cumulating the random index vectors of the words in its context, in an online fashion.

When applied to text classification [Sahlgren and Cöster, 2004], RI is reformulated to build distributional vectors for documents (and not for words), considering the entire document as the context. That is, documents end up being represented by the aggregation of the random index vectors of the terms they contain (weighted by their *tfidf* score). RI has been tested in (monolingual) text classification using SVMs as the classifier [Sahlgren and Cöster, 2004] but, to the best of our knowledge, has never been applied to poly-lingual contexts.

RI depends on two parameters: k, the percentage of non-zero values in the random index vectors, and n, the predefined dimensionality of the mapping. Setting $k = 1\%n$ is a common practice in the related literature (denoted $RI_{1\%}$ in our experiments), since it is known to help in preserving matrix sparsity.

Beside sparsity, the choice of k in RI has an effect on the probability that two random index vectors are orthogonal. We observed that the probability or orthogonality, i.e., $\langle \vec{u}, \vec{v} \rangle = 0$, for any two randomly generated vectors $\vec{u}, \vec{v}$ tends to increase as k decreases (it also increases when n increases, but by a fairly slower ratio – a broader discussion is later offered). Additionally, it could be shown that even for small values of k it is possible to encode a large number of distinct features in a reduced space (provided n is sufficiently high). As an example, by setting $k = 2$ and $n = 5,000$ it is possible to encode $2n(n-1) = 49,990,000$ distinct features. Note that this capacity already exceeds the demands of big datasets, while it still allows for a drastic reduction in the original feature space (typically, of the order of hundreds of thousands of dimensions).

Accordingly, and contrarily to the common $RI_{1\%}$ setting, we propose to set $k = 2$, i.e., to the smallest possible value for which a RI projection is still feasible[1]. We dub this configuration *Lightweight Random Indexing* (LRI). LRI presents the following advantages with respect to standard RI$_{1\%}$ and, in general, with respect to any RI with $k > 2$:

- Each index vector has only two non-zero values. The mapping can be allocated in memory for any number of original features, and the projection is performed very quickly.
- Given a fixed value of n, LRI has a higher probability of generating nearly-orthogonal projections.
- Parameter k becomes a constant that needs no tuning.

3 Experiments

We have compared LRI against (1) the polylingual BoW representation (PolyBoW – a language-agnostic BoW representation which simply represents all documents in the same feature space where each distinct term, irrespective of its language, is given a dedicated dimension), (2) the naïve monolingual (MonoBoW) classifiers[2] [García Adeva *et al.*, 2005], (3) Random Indexing with $k = n/100$ (RI$_{1\%}$) [Sahlgren and Cöster, 2004], (4) the Achlioptas mapping [Achlioptas, 2001], (5) Cross-Lingual Latent Semantic Analysis (CL-LSA) [Dumais *et al.*, 1997], (6) Multilingual Domain Models (MDM) [Gliozzo and Strapparava, 2005], and (7) feature selection on PolyBoW (FS)[3].

We have run experiments on different polylingual tasks, including classification of polylingual comparable documents, dimensionality reduction of the polylingual feature space, and the improvement of different monolingual classifiers by leveraging poly-lingual information[4]. The classifier was, in all cases, generated via SVMs with a linear kernel and with the rest of the parameters left to their default values.

As the effectiveness measures we use the macro- and micro-averaged versions of F_1, the harmonic mean of precision (π) and recall (ρ).

As the datasets on which to test our model we have considered two publicly available multilingual corpora, the Reuters RCV1/RCV2 and the JRC-Acquis collections. For the sake of brevity, in this extended abstract we will restrict our attention to the RCV1/RCV2 experiments, and refer the interested reader to [Moreo Fernández *et al.*, 2016] for more details.

RCV1/RCV2[5] consists of 804,414 English news stories (RCV1) plus 487,000 news stories written in other thirteen languages (RCV2) and produced by Reuters from 20 Aug 1996 to 19 Aug 1997. The collection is comparable at topic level, i.e., news stories are not direct translations of each other but are instead news referring to similar or related events written by local reporters in different languages. We randomly selected 8,000 news stories for 5 languages (English, Italian, Spanish, French, German) pertaining to the last 4 months (from 1997-04-19 to 1997-08-19), and we performed a 70%/30% train/test split, thus obtaining a training set of 28,000 documents (5,600 for each language) and a test set of 12,000 documents (2,400 for each language). In our experiments we have restricted our attention to the 67 classes (out of 103) with at least one positive training example for each of the five languages. After preprocessing (stopword removal

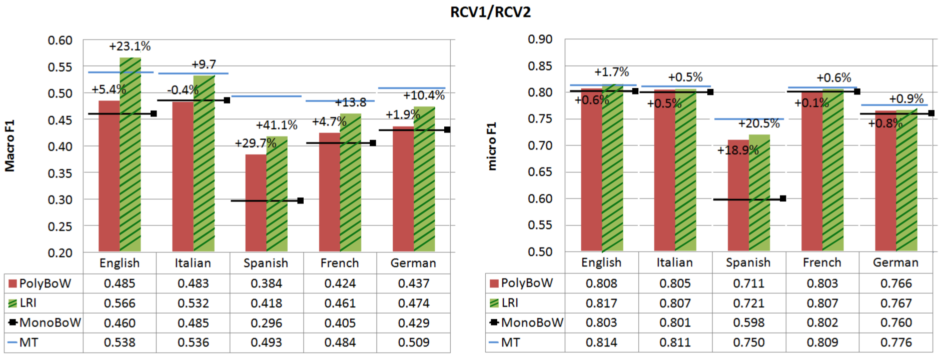
[1] Note that choosing $k = 1$ is equivalent to performing a random permutation of feature indexes in a bag-of-words (BoW) representation when $n = |F|$, and is insufficient to distinctly encode all features if $n < |F|$.

[2] PolyBoW and MonoBoW correspond to the NP1C and NPNC setups in [García Adeva *et al.*, 2005].

[3] Information Gain was used as the term space reduction function and round robin as the selection policy.

[4] The code all experiments is publicly available at http://hlt.isti.cnr.it/jatecs/ as part of the JaTeCs project [Esuli *et al.*, 2017]

[5] http://trec.nist.gov/data/reuters/reuters.html



	English	Italian	Spanish	French	German
🟥 PolyBoW	0.485	0.483	0.384	0.424	0.437
▨ LRI	0.566	0.532	0.418	0.461	0.474
■ MonoBoW	0.460	0.485	0.296	0.405	0.429
— MT	0.538	0.536	0.493	0.484	0.509

	English	Italian	Spanish	French	German
🟥 PolyBoW	0.808	0.805	0.711	0.803	0.766
▨ LRI	0.817	0.807	0.721	0.807	0.767
■ MonoBoW	0.803	0.801	0.598	0.802	0.760
— MT	0.814	0.811	0.750	0.809	0.776

Figure 1: Monolingual classification on RCV1/RCV2 using F_1^M (left) and F_1^μ (right) as the evaluation measures.

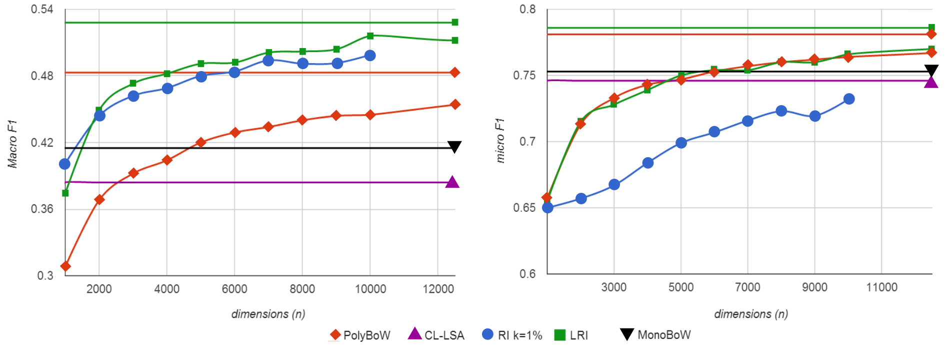


Figure 2: Dimensionality reduction experiments on RCV2/RCV1.

and stemming) the number of distinct features amounted to 123,258.

3.1 Results

Polylingual Information

As a first case of study, we investigate how much the addition of polylingual information affects the performance of a monolingual classifier. In this scenario we compare MonoBoW, where training is performed only on documents of the same language of the test documents; PolyBoW, which trains on documents from all the five languages; and LRI, for which we set $n = |F|$ (i.e., no dimensionality reduction).

The results shown in Figure 1 show that the simple addition of examples in different languages (PolyBoW) helps to improve the performance for the monolingual task, probably because of the shared words across languages; however, LRI clearly outperforms both MonoBoW and PolyBoW. The improvements are more marked for F_1^M than for F_1^μ, indicating that the improvements especially take place in the more infrequent classes, which impact susbstantially on F_1^M but not on F_1^μ.

Note that in this experiment the matrices that PolyBoW and LRI feed to the learning algorithm are of the same size. However, in LRI all dimensions become potentially useful for all languages due to random projection.

Dimensionality Reduction

The following experiments explore the dimensionality reduction aspect of the problem, and concern a realistic polylingual scenario, where both training and test data are of a multilingual nature. To test the scalability of LRI when several languages are involved we conducted an experiment varying the feature space dimension from 500 to 10,000. Due to its high computational costs, we varied the dimensionality of CL-LSA from 500 to 1000 (Figure 2); for MDM the dimensionality was set to 400, according to indications reported in [Gliozzo and Strapparava, 2005]. Note that not all algorithms were able to complete their execution due to memory constraints.

LRI obtained good results on both macro- and micro-averaged F_1, while the other methods exhibited alternating performance on the two measures. $\text{RI}_{1\%}$ obtained comparable results in terms of F_1^M but performed poorly on F_1^μ; in contrast, PolyBoW performed comparably in terms of F_1^μ but worse in terms of F_1^M. Surprisingly CL-LSA and MDM performed worse than the naïve classifier (MonoBoW) with all features. However, it should be remarked that they outperformed all other baselines in F_1^μ with only 400 (MDM) and 1000 (CL-LSA) dimensions.

4 Analysis: Space and Time Efficiency

The random projection has a direct impact on sparsity. Each time a document contains a feature, k non-zero values are placed in the projected matrix. It is usually the case that sparsity benefits not only space occupation but also execution time. To explore this phenomenon we run experiments considering only English and Italian documents, which generate

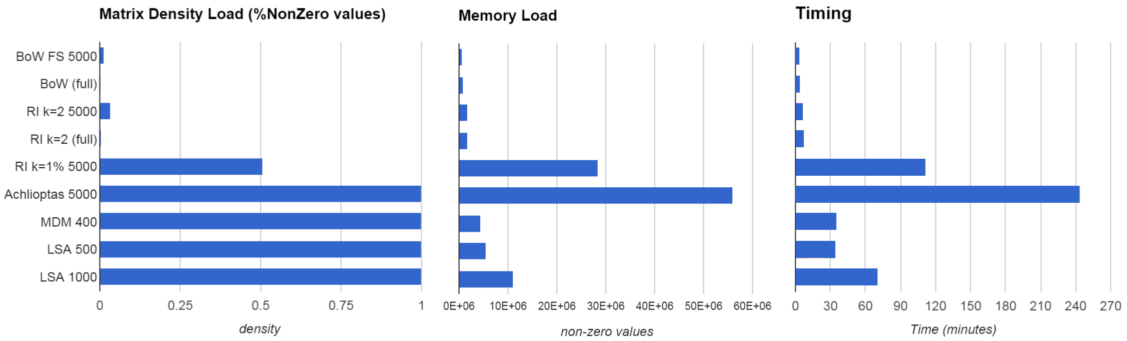


Figure 3: Reuters RCV1/RCV2 English and Italian (11,200-by-51,828 full training matrix size). Left: matrix density; center: memory load; right: execution time.

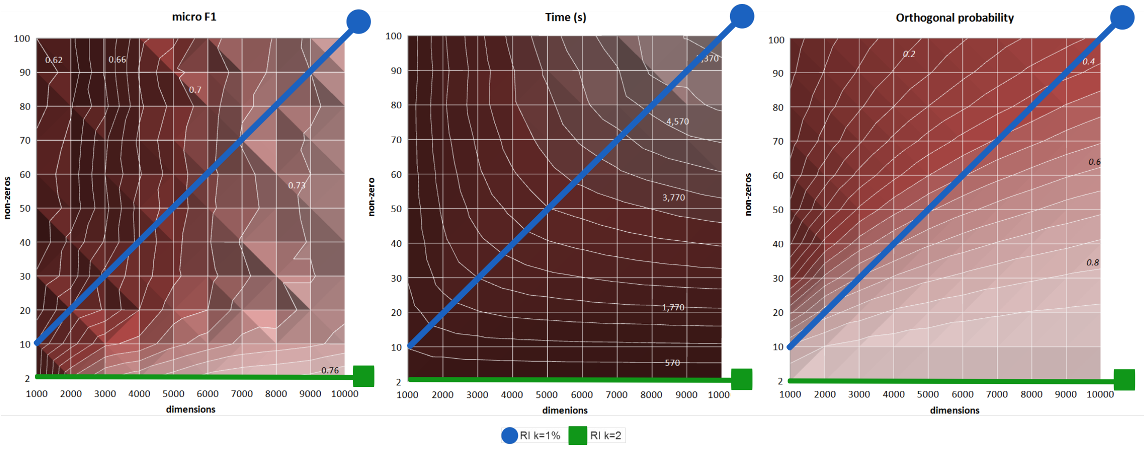


Figure 4: Contour lines on the impact of dimensionality n and non-zero values k on performance (left), execution time (center), and probability of finding an orthogonal pair of random indexes (right). Darker regions represent lower values.

a 11,200-by-51,828 co-occurrence matrix, and examine their matrix density load (percentage of non-zero values over the total matrix size), memory load (absolute number of non-zero values), and execution time[6] (Figure 3). LRI requires double the space of standard BoW, but succeeds to preserve sparsity, while $\text{RI}_{1\%}$ drastically increases the matrix density and produces a large memory load. MDM, LSA, and Achlioptas' method operate on dense matrices, which has a clear impact in execution times. The total time needed by LRI is roughly higher by a factor of 2 with respect to the full BoW representation, which is still negligible if compared with the rest of baselines, especially $\text{RI}_{1\%}$ and the Achlioptas mapping.

We empirically studied the probability function of this event as a function of the number of non-zero values (k) for RI, ranging from 2 to 100, and of the reduced dimensionality (n), ranging from 1,000 to 10,000 (Figure 4).

The following trends can be directly observed from the results. The performance in $\text{RI}_{1\%}$ improves at the cost of space and time efficiency, and by gradually disrupting the orthogonality of the base. On the contrary, LRI behaves differently: when dimensionality increases, (i) accuracy improves (ii) without penalizing execution times due to the preservation of sparsity, and (iii) the orthogonality of the base is improved.

5 Conclusions

Lightweight Random Indexing is a machine-translation-free, dictionary-free variant of RI that better preserves matrix sparsity (i.e., both memory load and training times are not penalized) and increases the chance of orthogonality among random vectors. This configuration yielded the best results in classification accuracy in two popular multilingual benchmarks.

In the polylingual BoW representation most of the features (dimensions) are only informative for one of the languages. RI instead maps the feature space into a space that is shared among all languages at once. The effect is that any dimension of the space becomes informative regardless of language. In a particular configuration in which the projection space is larger than the actual number of different features for a single language, the “kernel-trick” effect appears: the informative space for each language is enlarged and it thus becomes more easily separable.

[6] All the experiments were run on a dedicated Intel i7 64bit processor with 12 cores, running at 1,600MHz, and 24GBs RAM memory.

References

[Achlioptas, 2001] Dimitris Achlioptas. Database-friendly random projections. In *Proceedings of the 20th ACM Symposium on Principles of Database Systems (PODS 2001)*, pages 274–281, Santa Barbara, US, 2001.

[Amini *et al.*, 2009] Massih-Reza Amini, Nicolas Usunier, and Cyril Goutte. Learning from multiple partially observed views; An application to multilingual text categorization. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS 2009)*, pages 28–36, Vancouver, CA, 2009.

[Bel *et al.*, 2003] Nuria Bel, Cornelis H. Koster, and Marta Villegas. Cross-lingual text categorization. In *Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2003)*, pages 126–139, Trondheim, NO, 2003.

[Deerwester *et al.*, 1990] Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[Dumais *et al.*, 1997] Susan T. Dumais, Todd A. Letsche, Michael L. Littman, and Thomas K. Landauer. Automatic cross-language retrieval using latent semantic indexing. In *Working Notes of the AAAI Spring Symposium on Cross-language Text and Speech Retrieval*, pages 18–24, Stanford, US, 1997.

[Esuli *et al.*, 2017] Andrea Esuli, Tiziano Fagni, and Alejandro Moreo Fernández. Jatecs an open-source java text categorization system. *arXiv preprint arXiv:1706.06802*, 2017.

[Fradkin and Madigan, 2003] Dmitriy Fradkin and David Madigan. Experiments with random projections for machine learning. In *Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2003)*, pages 517–522, Washington, US, 2003.

[García Adeva *et al.*, 2005] Juan José García Adeva, Rafael A. Calvo, and Diego López de Ipiña. Multilingual approaches to text categorisation. *European Journal for the Informatics Professional*, 5(3):43–51, 2005.

[Gliozzo and Strapparava, 2005] Alfio Gliozzo and Carlo Strapparava. Cross-language text categorization by acquiring multilingual domain models from comparable corpora. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 9–16, Ann Arbor, US, 2005.

[Johnson *et al.*, 1986] William B. Johnson, Joram Lindenstrauss, and Gideon Schechtman. Extensions of Lipschitz maps into Banach spaces. *Israel Journal of Mathematics*, 54(2):129–138, 1986.

[Kanerva *et al.*, 2000] Pentti Kanerva, Jan Kristofersson, and Anders Holst. Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 1036–1037, Philadelphia, US, 2000.

[Moreo Fernández *et al.*, 2016] Alejandro Moreo Fernández, Andrea Esuli, and Fabrizio Sebastiani. Lightweight random indexing for polylingual text classification. *Journal of Artificial Intelligence Research*, 57:151–185, 2016.

[Nastase and Strapparava, 2013] Vivi Nastase and Carlo Strapparava. Bridging languages through etymology: The case of cross-language text categorization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 651–659, Sofia, BL, 2013.

[Papadimitriou *et al.*, 1998] Christos H. Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. In *Proceedings of the 17th ACM Symposium on Principles of Database Systems (PODS 1998)*, pages 159–168, Seattle, US, 1998.

[Sahlgren and Cöster, 2004] Magnus Sahlgren and Rickard Cöster. Using bag-of-concepts to improve the performance of support vector machines in text categorization. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, CH, 2004.

[Sahlgren and Karlgren, 2005] Magnus Sahlgren and Jussi Karlgren. Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering*, 11(3):327–341, 2005.

[Sahlgren, 2001] Magnus Sahlgren. Vector-based semantic analysis: Representing word meanings based on random labels. In *Proceedings of the ESSLLI Workshop on Semantic Knowledge Acquistion and Categorization*, Helsinki, FI, 2001.

[Sahlgren, 2005] Magnus Sahlgren. An introduction to random indexing. In *Proceedings of the Workshop on Methods and Applications of Semantic Indexing*, Copenhagen, DK, 2005.

[Vinokourov *et al.*, 2002] Alexei Vinokourov, John Shawe-Taylor, and Nello Cristianini. Inferring a semantic representation of text via cross-language correlation analysis. In *Proceedings of the 16th Annual Conference on Neural Information Processing Systems (NIPS 2002)*, pages 1473–1480, Vancouver, CA, 2002.

[Wei *et al.*, 2011] Chih-Ping Wei, Yen-Ting Lin, and Christopher C. Yang. Cross-lingual text categorization: Conquering language boundaries in globalized environments. *Information Processing and Management*, 47(5):786–804, 2011.

[Wei *et al.*, 2014] Chih-Ping Wei, Chin-Sheng Yang, Ching-Hsien Lee, Huihua Shi, and Christopher C. Yang. Exploiting poly-lingual documents for improving text categorization effectiveness. *Decision Support Systems*, 57:64–76, 2014.

[Xiao and Guo, 2013] Min Xiao and Yuhong Guo. A novel two-step method for cross-language representation learning. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS 2013)*, pages 1259–1267, Lake Tahoe, US, 2013.