Lightweight Random Indexing for Polylingual Text Classification

by Alejandro Moreo Fernandez, Andrea Esuli and Fabrizio Sebastiani (ISTI-CNR)

Researchers from ISTI-CNR, Pisa (in a joint effort with the Qatar Computing Research Institute), have undertaken an effort aimed at producing more accurate and more efficient means of performing poly-lingual text classification, i.e., automatic text classification in which classifying text in one language can also leverage training data expressed in a different language.

Multilingual Text Classification (MLTC) is a text classification task in which documents are written each in one among a set L of natural languages, and in which all documents must be classified under the same classification scheme, irrespective of language. This scenario is more and more frequent, given the large quantity of multilingual platforms and communities emerging on the Internet.

There are two main variants of MLTC, namely Cross-Lingual Text Classification (CLTC) and Polylingual Text Classification (PLTC). In CLTC we assume that for one or more of the languages in L there are no training documents; the task thus consists of classifying the test documents expressed in these languages by leveraging the training documents expressed in the other languages. In PLTC, which is the focus of this work, we assume (differently from CLTC) that for each language in L there is a representative set of training documents; PLTC consists of improving the accuracy of each of the |L| monolingual classifiers by also leveraging the training documents written in the other (|L|-1) languages. This task is receiving increased attention in the text classification community also due the new challenge it poses, i.e., how to effectively leverage polylingual resources in order to infer a multilingual classifier and to improve the performance of a monolingual one.

The obvious solution, consisting of generating a single polylingual classifier from the juxtaposed monolingual vector spaces, is usually infeasible, since the dimensionality of the resulting vector space is roughly |L| times that of a monolingual one, and is thus often unmanageable. As a response, the use of machine translation tools or multilingual dictionaries has been proposed. However, these resources are not always available, or are not always free to use.

One machine-translation-free and dictionary-free method that had never been applied to PLTC before, is Random Indexing (RI). RI is a context-counting model belonging to the family of random projection methods, which produces linear projections into a nearly-orthogonal reduced space where the original distances between vectors are approximately preserved. RI thus delivers semantically meaningful representations in a reduced space, and can be viewed as a cheaper approximation of Latent Semantic Analysis. We have analysed RI in terms of space and time efficiency, and have proposed as a result a particular configuration of it (that we have dubbed Lightweight Random Indexing -- LRI). LRI is designed so that the orthogonality of the projection base is maximized, which causes sparsity to be preserved after the projection (see Figure 1). The orthogonality of random index vectors plays an important role for features that are shared across languages: if their corresponding random index vectors are orthogonal with respect to all the other vectors, the information they contribute to the process is maximized, instead of being diluted by other less informative features.

We have run experiments on two well-known public benchmarks, Reuters RCV1/RCV2 (a comparable corpus -- i.e., documents are not direct translations of each other, but are simply about similar topics) and JRC-Acquis (a parallel one -- i.e., each text is available in all languages thanks to the intervention of human translators); for both benchmarks, we addressed five languages (English, Italian, Spanish, French, German). These experiments have shown LRI to outperform (both in terms of effectiveness and efficiency) a number of previously proposed machine-translation-free and dictionary-free PLTC methods that we used as baselines, including other more classical instantiations of random indexing.

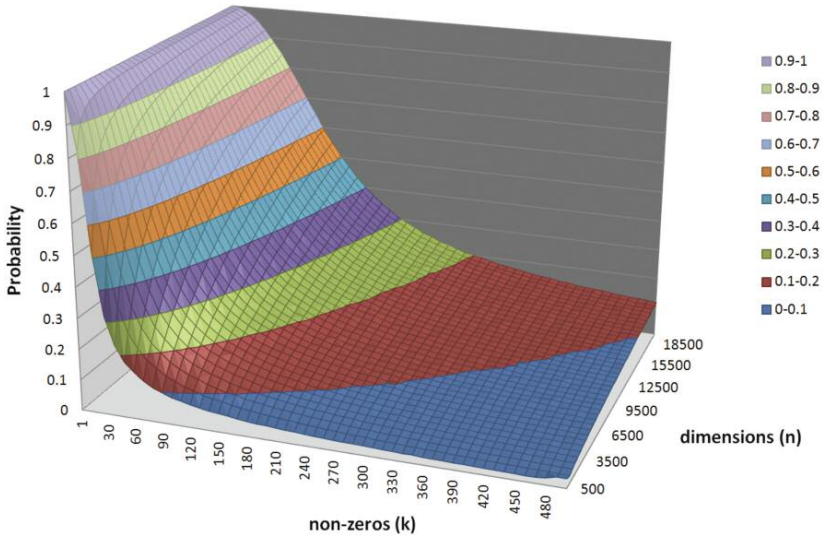

Figure 1: Variation in the probability of orthogonality of two random index vectors as a function the number of non-zero values in the random index vector and its dimensionality.

Link:
http://http://jair.org/papers/paper5194.html

Please contact:
Fabrizio Sebastiani, ISTI-CNR, Italy
+39 050 6212892, fabrizio.sebastiani@isti.cnr.it