On Dimensionality Reduction Techniques for Cross-Language Information Retrieval

Parth Gupta
Natural Language Engineering Lab - ELiRF
Department of Information Systems and Computation
Universidad Politecnica de Valencia, Spain
`http://www.dsic.upv.es/grupos/nle`
pgupta@dsic.upv.es

With the advent of the Web, cross-language information retrieval (CLIR) becomes important not only to satisfy the information need across languages but to *mine* resources for multiple languages e.g. parallel or comparable documents. Broadly CLIR techniques are of two types, in the first case, either queries or documents are translated to the language of comparison while the other type tries to project the vector space representation of the text to a shared translingual space which represents the "semantics" of the documents. In this study, we review the state-of-the-art for CLIR by means of the latter approach and identify the scope for further research.

CLIR, IR, dimensionality reduction, non-linear

1. INTRODUCTION

The most celebrated approach of dimensionality reduction for CLIR is cross-language latent semantic indexing (CL-LSI) Dumais *et al.* (1997). CL-LSI falls into the category of linear dimensionality reduction techniques which by using parallel documents tries to reduce the dimensionality to top k principal components to represent the data in reduced "semantic" space. There has been an extension to CL-LSI called oriented principal component analysis (OPCA) which formulates the problem as generalised eigenproblem Platt *et al.* (2010).

Under non-linear approaches to dimensionality reduction, cross-language multidimensional scaling (CL-MDS) was used to project the documents across languages in Banchs and Kaltenbrunner (2008) in reduced space. It shows how to exploit the structural information through monolingual learning and then projecting one language document in the space of the other languages. The linear approaches like PCA embed the data into low-dimensional hyperplane. The problem with CL-MDS is, it being a transductive method, does not have an operator matrix like projection matrix of CL-LSI. Hence, it does not generalise for the unseen data which is a big restriction for IR systems.

When data components have non-linear dependencies, linear approaches often require larger dimensions compared to the non-liner counterparts. Recently, Hinton and Salakhutdinov (2006) showed that how efficiently and effectively a deep architecture of neural networks can be trained to induce more abstract representation of the data with much smaller dimensions compared to PCA (30 vs. 128). Salakhutdinov and Hinton (2007) show the application of this framework to IR.

2. TRANSLINGUAL DIMENSIONALITY REDUCTION

In this section we review a few variety of dimensionality reduction techniques for CLIR. Commonly all of them represent the data in vector space form as a matrix D of the training collection C with n documents in each language. This collection is usually parallel or comparable but for the further discussion we will consider only parallel, where document $d^1_{l_1}$ and $d^1_{l_2}$ are the translations of each other. The dimension of document-term matrix D is $n \times k$ where k is the vocabulary size including both languages. Each row i of D specifies a pair of parallel documents and D_{ij} is the TF-IDF score of term j in document $d^i_{l_1}$ or $d^i_{l_2}$ where $d^i_{l_1}$ and $d^i_{l_2}$ are i^{th} parallel documents in language l_1 and l_2 respectively. The dimensionality reduction techniques mostly differ in the ways they formulate D and solve it to find reduced dimensions.

2.1. Cross-Language Latent Semantic Indexing (CL-LSI)

CL-LSI basically performs singular value decomposition of D in the lines of principal component analysis (PCA) (Dumais *et al.* 1997). CL-LSI obtains top k principal components of D which is considered as projection space and documents are compared in this space. The inherent idea is semantically similar terms across languages (dimensions of D) will correspond to similar latent components and these documents are near to each other in the reduced comparison space.

This method can also be looked as eigenproblem which is formulated as below:

$$Cv_j = \lambda_j v_j, \tag{1}$$

where, λ_j is the j^{th} largest eigenvalue, v_j is corresponding eigenvector and C is correlation matrix ($D^T D$). CL-LSI uses top k eigenvectors for projection.

2.2. Oriented Principal Component Analysis (OPCA)

OPCA formulates the problem as generalised eigenproblem which in the essence maximises the signal-to-noise ratio (Platt *et al.* 2010).

$$Sv_j = \lambda_j N v_j, \tag{2}$$

where, S is C-like matrix and N is covariance matrix of the differences among parallel documents which is considered as noise.

Theoretically, OPCA tries to minimise the distance between the parallel documents at the same time maximising the overall variance of the data.

2.3. Cross-Language Multidimensional Scaling (CL-MDS)

CL-MDS uses non-linear dimensionality reduction technique called MDS to model monolingual reduced dimensional maps of parallel collection (Banchs and Kaltenbrunner 2008). Then, it uses the structural similarity of these maps to CLIR. Primarily, it prepares monolingual maps of each language using non-linear MDS projection through vector space D-like matrix of each language independently. These documents are called anchor documents for each language. New documents are projected in the other language's space through a transformation matrix T,

$$T = MD^{-1}, \tag{3}$$

where, D is an $n \times n$ matrix of cosine distances between the anchor documents in the original space and M is $k \times n$ matrix of k dimensions of n anchor documents in then projected space. Finally, a document in l_1 can be placed in the map of l_2 by computing distances with respect to l_1 anchor documents in the original vector space and using the transformation matrix T computed with anchor documents in l_2.

2.4. Cross-Language Deep Belief Networks (CL-DBN)

CL-DBN builds deep belief network for each language in deep framework where each layer is Restricted Boltzmann Machines (RBM) (Kim *et al.* 2012). These deep networks are trained through greedy layer-by-layer pretraining followed by the supervised fine-tuning. The structure of the network and the training procedure is shown in Fig 1. For more details on training and structure see (Kim *et al.* 2012). The dimensions of the projection space of each language are mapped using canonical correlation analysis (CCA) and similarity is estimated in this space.

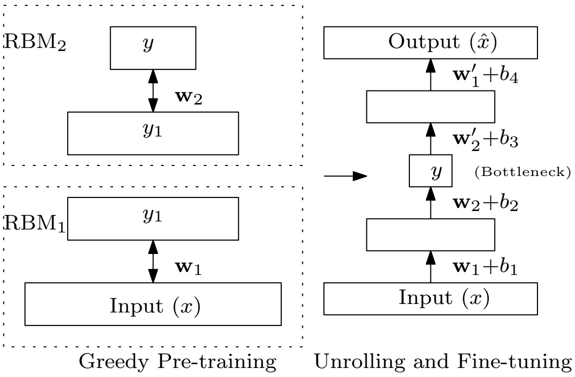

Figure 1: ***Left panel****: pre-training the stacked RBMs where upper RBMs take output of the lower RBM.* ***Right panel****: After pre-training the structure is "unrolled" to create a multi-layer network which is fine-tuned by backpropagation to perform $\hat{x} \approx x$.*

3. RESEARCH PROBLEM

Non-linear approaches to dimensionality reduction have shown to extract better abstract level representation of text for "semantic" comparison (Hinton and Salakhutdinov 2006). CL-MDS is a transductive method and hence doesn't have an operator matrix to transform unseen documents into reduced space which makes it limited to the anchor documents. Hinton and Salakhutdinov (2006) clearly outlines the suitability of deep learning for learning abstract projections of text using it. We plan to further investigate the potential of deep learning for abstract projections of text across languages. There are attempts to link *deep* projections to facilitate CLIR e.g. (Kim *et al.* 2012), but we feel, it is still not properly investigated in depth. Deep learning for dimensionality reduction is non-trivial problem compared to other approaches like CL-LSI, OPCA or CL-MDS because it is a non-convex optimisation problem with many local-minima(Erhan *et al.* 2010). Learning can go wrong for n-number of reasons sometimes without explicit evidence. Rather most of the literature comprise of the end-application based evaluation, in which case, it becomes more difficult to judge if the learning went correct or not especially in case of bad results. So there is a scope to measure the quality of learning for the task at hand and after gaining the confidence to work out better dimension-mapping strategies.

REFERENCES

Banchs, R. E. and Kaltenbrunner, A. (2008). Exploiting MDS projections for Cross-language IR. In *SIGIR*, pages 863–864.

Dumais, S., Landauer, T. K., and Littman, M. L. (1997). Automatic Cross-Linguistic Information Retrieval using Latent Semantic Indexing.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., and Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *J. Mach. Learn. Res.*, **11**, 625–660.

Hinton, G. and Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science*, **313**(5786), 504 – 507.

Kim, J., Nam, J., and Gurevych, I. (2012). Learning semantics with deep belief network for cross-language information retrieval. In *COLING (Posters)*, pages 579–588.

Platt, J. C., Toutanova, K., and tau Yih, W. (2010). Translingual document representations from discriminative projections. In *EMNLP*, pages 251–261.

Salakhutdinov, R. and Hinton, G. (2007). Semantic Hashing. In *SIGIR workshop on Information Retrieval and applications of Graphical Models*.