Exploiting MDS Projections for Cross-language IR

Rafael E. Banchs
rafael.banchs@barcelonamedia.org

Andreas Kaltenbrunner
andreas.kaltenbrunner@barcelonamedia.org

Barcelona Media - Innovation Centre
Ocata 1, 08003 Barcelona, Spain
+34-93-542-1100

ABSTRACT

In this paper, we describe some preliminary work on using monolingual projections of document collections for performing cross-language information retrieval tasks. The proposed methodology uses multidimensional scaling for projecting the vector-space representations of a given multilingual document collection into spaces of lower dimensionality. An independent projection is computed for each different language, and the structural similarities of the resulting projections are exploited for information retrieval tasks.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information search and retrieval—*clustering, retrieval models, search process.*

General Terms

Algorithms, Experimentation

Keywords

Cross-language IR, Multidimensional Scaling.

1. INTRODUCTION

Recently, cross-language information retrieval (CLIR) has been gaining importance as more multilingual information has become available in the World Wide Web. According to [3], four different strategies are typically used for CLIR. One of them is based on cognates, for which untranslatable and/or similar terms (in case of close languages) are used for matching the query. The other three strategies are based on using available techniques for: translating the query, translating the documents, or computing an intermediate representation (interlingua) for both, query and documents.

The most prominent method in the interlingua category uses latent semantic indexing (LSI) to create a multilingual vector-space representation for parallel collections [2]. In the method proposed here, instead of using a multilingual vector-space representation, monolingual vector representations are used for documents in a given multilingual collection. Then, each representation is projected into a lower dimensional space using a non-linear projection method, more specifically MDS (multidimensional scaling) [1]. Then the CLIR task is accomplished by exploiting the structural similarities among all resulting projections.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

2. DATA PROJECTIONS WITH MDS

Figure 1 presents two-dimensional projections for three monolingual collections of TF-IDF normalized bag-of-words representations of the 66 books of "The Holy Bible". The languages considered are English, Chinese and Spanish, with vocabulary sizes of 7265, 6286 and 13632 respectively. Remarkable similarities among the three data projections can be observed in figure 1. These projections were computed by using MDS, based on cosine distance metrics and Sammon's nonlinear mapping criterion [1].

3. THE PROPOSED METHODOLOGY

The idea behind the proposed methodology is to exploit structural similarities observed among the different monolingual projections computed with MDS to identify possible correspondences among new multilingual documents.

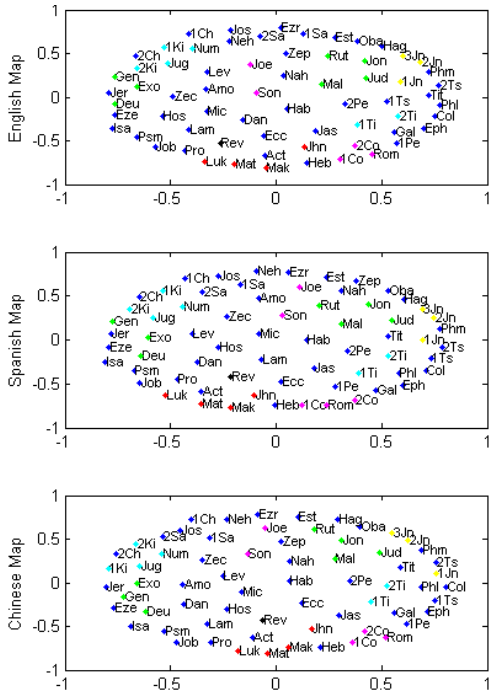

Figure 1: Two-dimensional projections for sixty six books of "The Holy Bible".

Suppose we are interested in using the projections of figure 1 for performing CLIR of new documents, any of the three monolingual maps can be actually used for the retrieval task. We will refer to the documents defining the maps as anchor documents. First of all, a procedure for placing any new document into its corresponding monolingual map is required (since MDS projections are data dependent, including the new document with the original anchor documents and performing a new projection with MDS is not an option). Two methods are proposed for document placement: a linear approximation, and optimization. Only the first method is evaluated here.

The linear projection approximation uses a transformation matrix T to project the distances between a given document vector d and all anchor documents measured in the original vector space into coordinates in the projected low-dimensional vector space. Such a transformation matrix is estimated from anchor documents as follows: $T = MD^{-1}$, where M is a $k \times n$ matrix containing the k coordinates of the n anchor documents in the projected vector space, and D is an $n \times n$ matrix containing the distances between every pair of anchor documents in the original vector space.

Optionally, an optimization procedure could be used to place a new document in the map preserving the ratios of its distances to all anchor documents as much as possible with respect to the distances in the original vector space.

Finally, CLIR can be achieved by using the described document placement methods to place documents of different languages in the same map. Any document in language A can be placed in a map in language B by computing distances with respect to language-A anchor documents in the original vector space and using the transformation matrix (or performing the optimization procedure) computed with language-B anchor documents.

4. EXPERIMENTAL RESULTS

In this section we present some preliminary results comparing the proposed methodology with the LSI approach [2].

Each result presented in figure 2 (continuous lines) represents the mean value of 30 retrieval experiments $\pm$ the corresponding standard deviation (dotted lines). For each experiment, 30 documents were randomly selected as training (anchor), and other 30 documents as test data collection.

The retrieval task consisted of recovering a document using the same document in a different language as a query. The $y-$axis represents the proportion of successful retrievals considering the top-1 retrieved document and the $x-$axis the space dimensionality used for the projections. Reduction of dimensionality to k in LSI is achieved by using only the largest k singular values. In our method k is a parameter of the MDS-projection and results were computed by placing all test documents into the English maps. Even for Spanish-Chinese CLIR, we used the English projection to place documents of both languages in the reduced space where the actual CLIR-task is performed.

5. CONCLUSIONS AND FUTURE WORK

A novel method for CLIR which exploits the structural similarity among MDS-based monolingual projections of a multilingual collection was proposed.

Preliminary results have shown that the performance of the proposed method is comparable to the LSI approach,

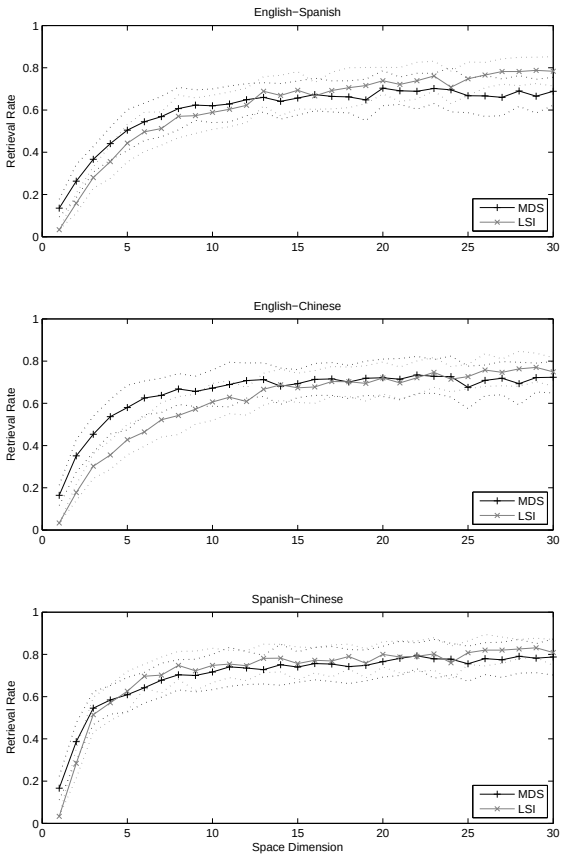

Figure 2: Performance for CLIR by using the proposed method (black) and LSI (gray) operating at different space dimensionalities.

but our method has the potentiality of exploiting all the available information existing in a typical multilingual collection. While LSI is restricted, by construction, to use only the intersection set of documents in a given multilingual collection (i.e. parallel documents available in all the languages of interest), our method can be trained using the union of the multilingual collections. E.g. we can use different subsets of the anchor documents in English for the projections of Spanish and Chinese documents and yet compare the results for proximity. Further experiments with larger datasets and more realistic queries are required to evaluate the practical implications of this theoretical advantage.

6. ACKNOWLEDGEMENTS

The authors want to thank Barcelona Media Innovation Centre for the support and permission to publish this work.

7. REFERENCES

[1] M. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman & Hall, London, UK, 2001.

[2] S. T. Dumais, T. K. Landauer, and M. L. Littman. Automatic cross-linguistic information retrieval using latent semantic indexing. In *SIGIR96 Workshop On Cross-Linguistic Information Retrieval*, 1996.

[3] K. Kishida. Technical issues of cross-language information retrieval: a review. *Information Processing and Management*, 41(3):433–455, 2005.