Multidimensional Scaling and cross-language information retrieval

Rafael E. Banchs

Barcelona Media
Centre d'Innovació

Multidimensional Scaling

A method for data visualization ***(Shepard,1962; Kruskal, 1964)***

- Given a set of similarity, dissimilarity or ordinal relations among a group of objects

- Find a set of Euclidean coordinates to the objects in the group (i.e. an embedding)

- Such that the relations obeyed by the objects are preserved as much as possible

Barcelona Media
Centre d'Innovació

Classes of MDS

Two main categories of multidimensional scaling are distinguished:

- *Metric multidimensional scaling:* Euclidean distances among the points in the embedding should match as much as possible the original object dissimilarities

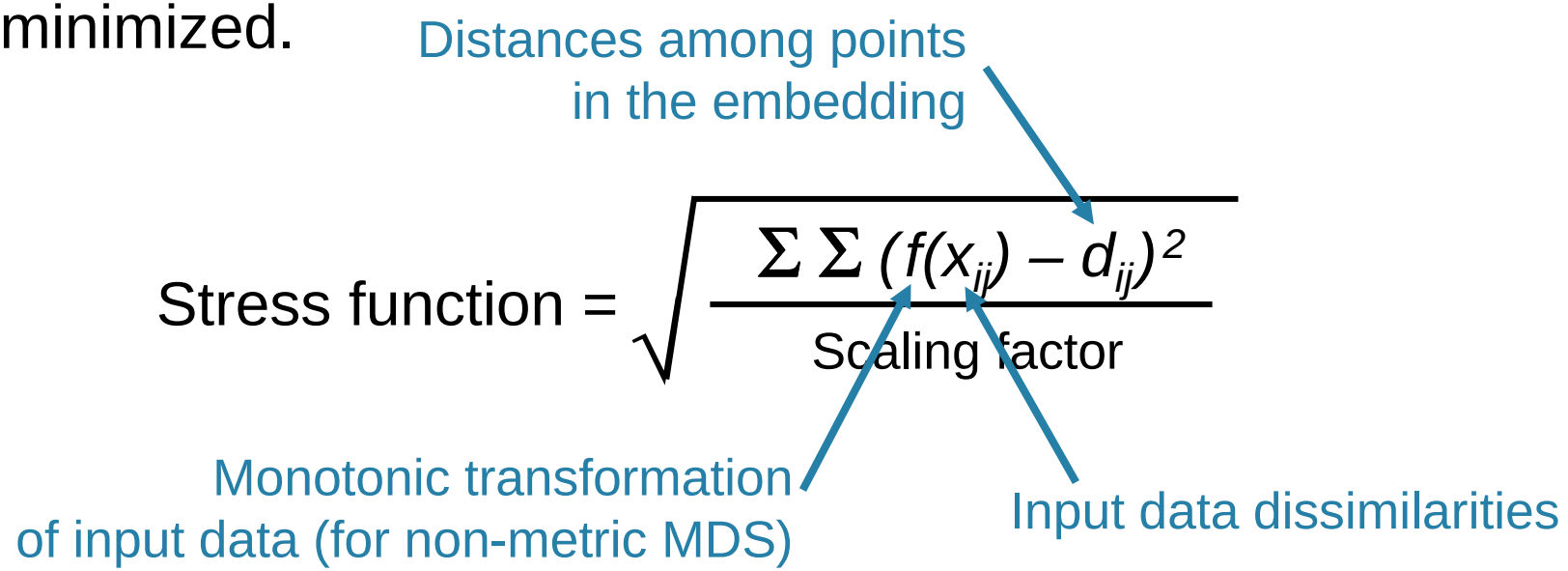
- *Non-metric multidimensional scaling:* only the relative ordering or ranks are attempted to be preserved (i.e. distances among points in the embedding approximate a *monotonic transformation* of original dissimilarities)

Barcelona Media
Centre d'Innovació

An optimization problem

MDS is implemented via an optimization problem:

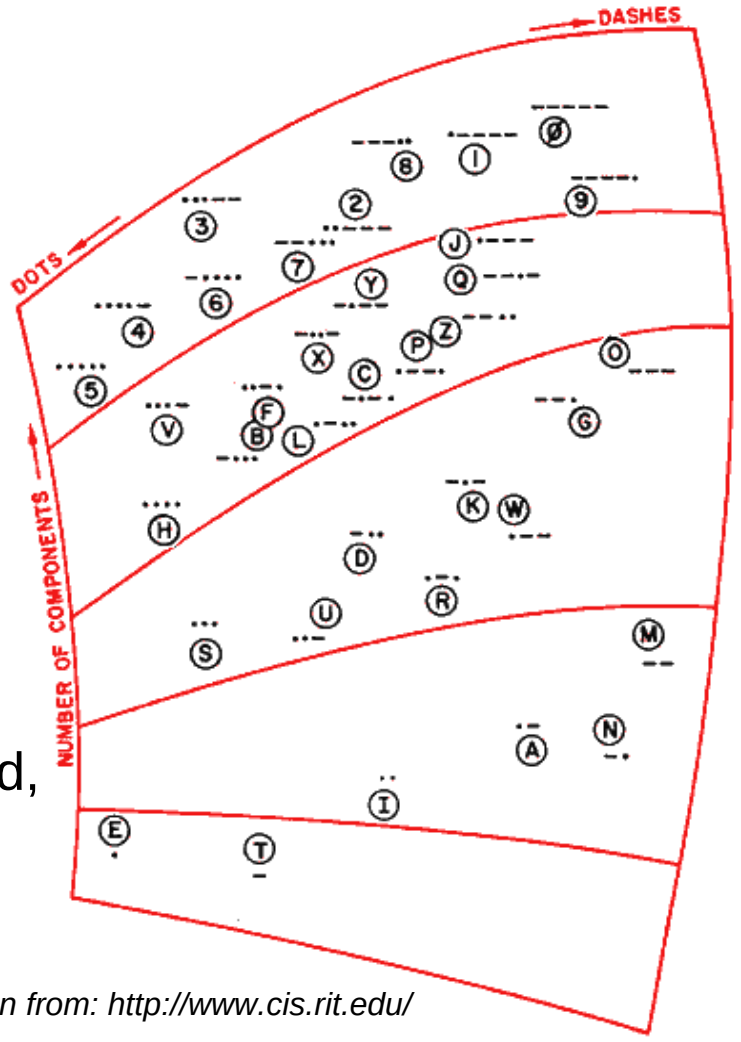
- Euclidean distances among all pairs of points in the embedding are adjusted such that a stress function is minimized.



Classical examples of MDS

A map for Morse code:

- 150 subjects (unfamiliar with the code) were presented with pairs of letters in Morse code
- the task was to say whether the two signals were the same or different
- the data formed a “similarity” matrix (the more times signals were confused, the greater the number in the cell that represented that pair of signals)



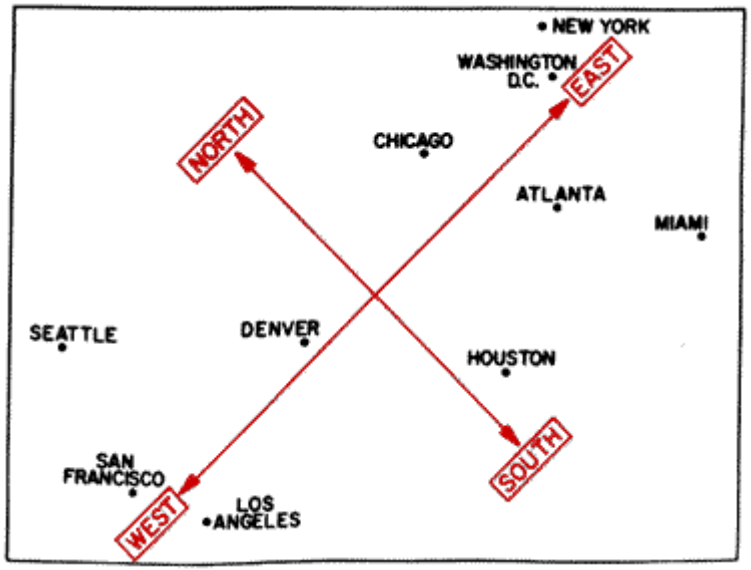
Taken from: http://www.cis.rit.edu/

Classical examples of MDS

Map generation from inter-city distances:

- For a given a matrix containing distances among cities, a map can be inferred…

CITIES	ATLA.	CHIC.	DENV.	HOUS.	L.A.	MIAMI	N.Y.	S.F.	SEAT.	WASH D.C
ATLANTA		587	1212	701	1936	604	748	2139	2182	543
CHICAGO	587		920	940	1745	1188	713	1858	1737	597
DENVER	1212	920		879	831	1726	1631	949	1021	1494
HOUSTON	701	940	879		1374	968	1420	1645	1891	1220
LOS ANGELES	1936	1745	831	1374		2339	2451	347	959	2300
MIAMI	604	1188	1726	968	2339		1092	2594	2734	923
NEW YORK	748	713	1631	1420	2451	1092		2571	2408	205
SAN FRANCISCO	2139	1858	949	1645	347	2594	2571		678	2442
SEATTLE	2182	1737	1021	1891	959	2734	2408	678		2329
WASHINGTON D.C.	543	597	1494	1220	2300	923	205	2442	2329	



Taken from: http://www.cis.rit.edu/

Barcelona Media
Centre d'Innovació

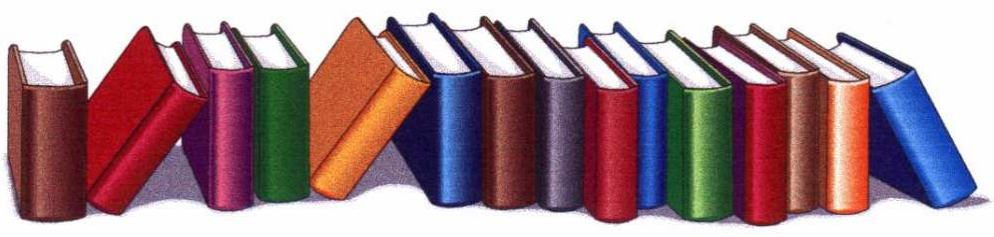
MDS applied to text data

Consider some document collection:

- Obtain a vector-space representation by using standard TFIDF weighting
- Compute dissimilarities among books by using cosine distances
- Construct a low dimensionality map of the collection by using MDS

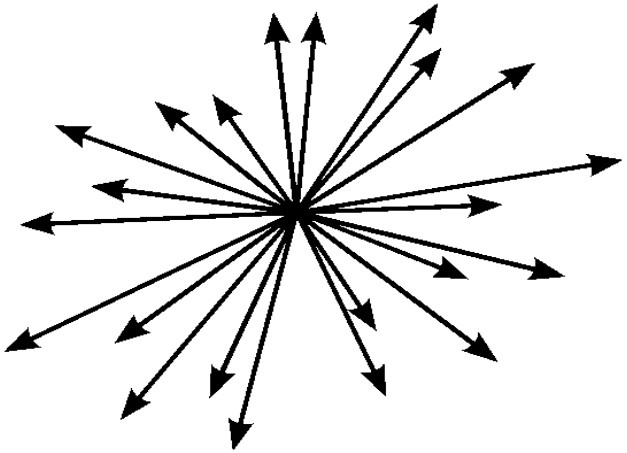
Barcelona Media
Centre d'Innovació

MDS applied to text data



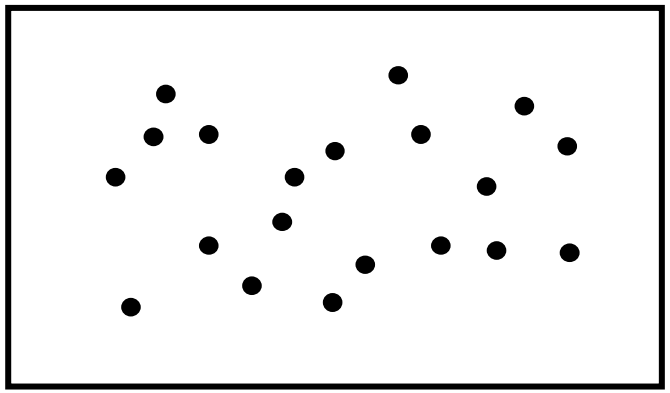
Document collection

TF-IDF



Vector-space representation

MDS



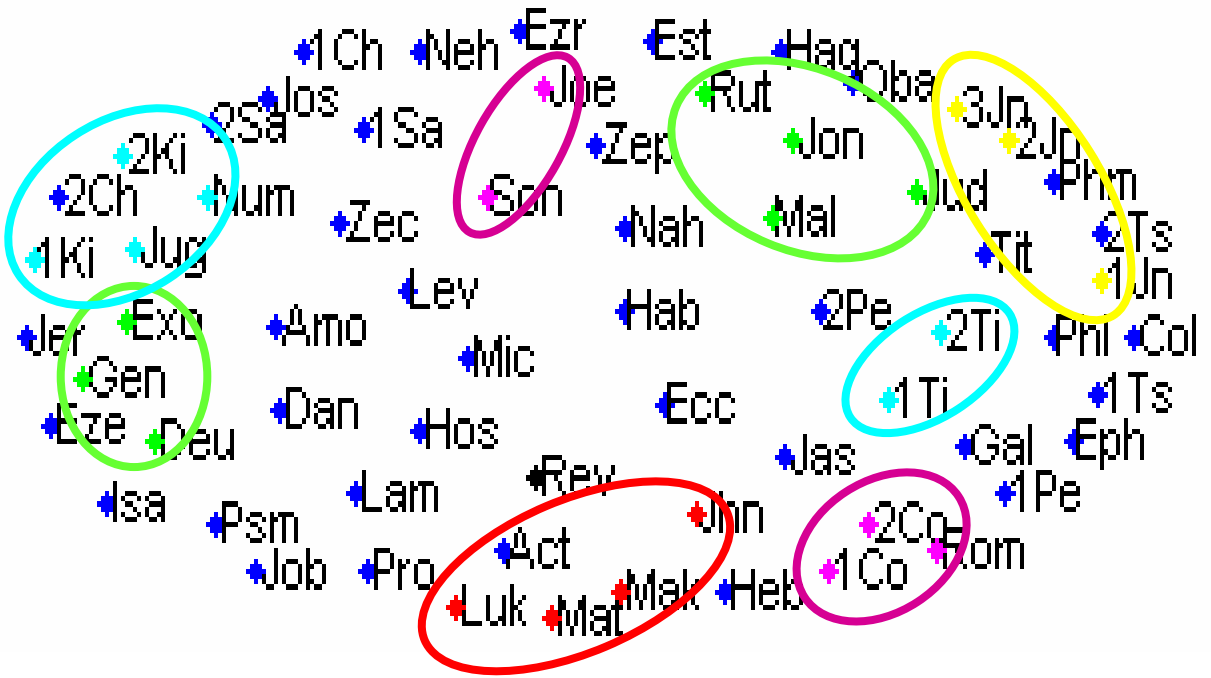
Low-dimensional map

Barcelona Media
Centre d'Innovació

A trilingual data collection

Consider 66 books from the “The Holy Bible”

- Chinese version (vocabulary size: 12952 words)

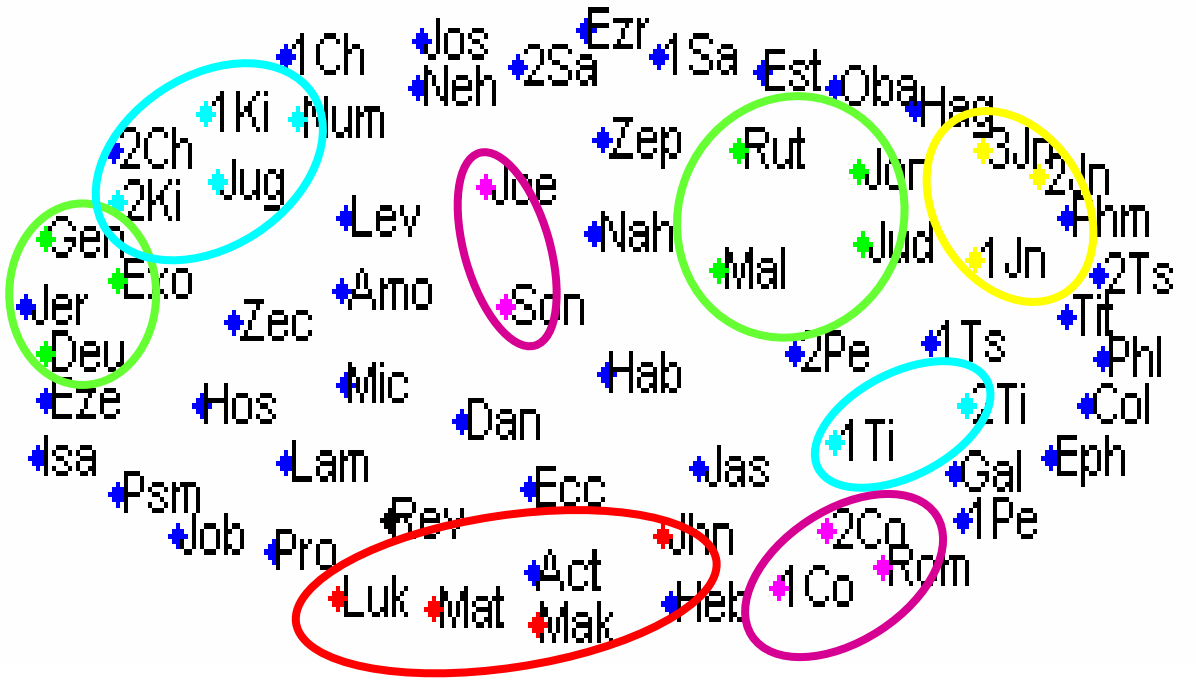


Barcelona Media
Centre d'Innovació

A trilingual data collection

Consider 66 books from the “The Holy Bible”

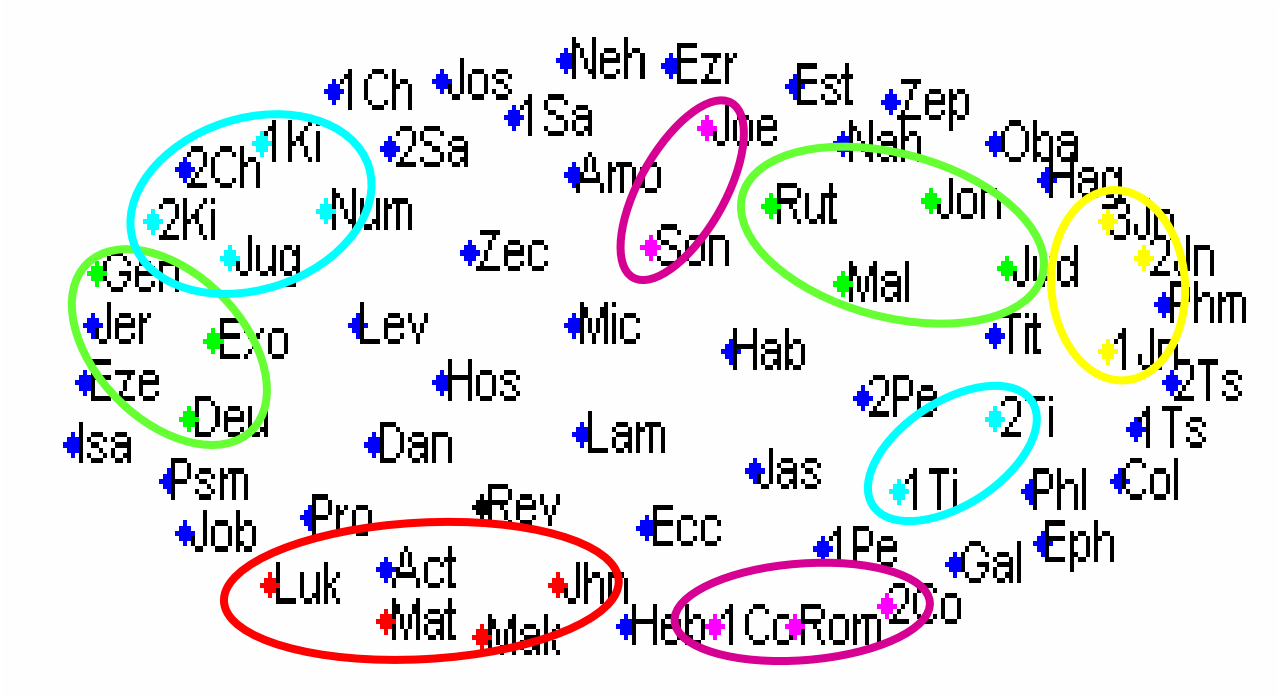
- English version (vocabulary size: 8121 words)



A trilingual data collection

Consider 66 books from the “The Holy Bible”

- Spanish version (vocabulary size: 25385 words)



Barcelona Media
Centre d'Innovació

Important observations

- The maps have been obtained for each language independently from each other language (i.e. in a monolingual context)

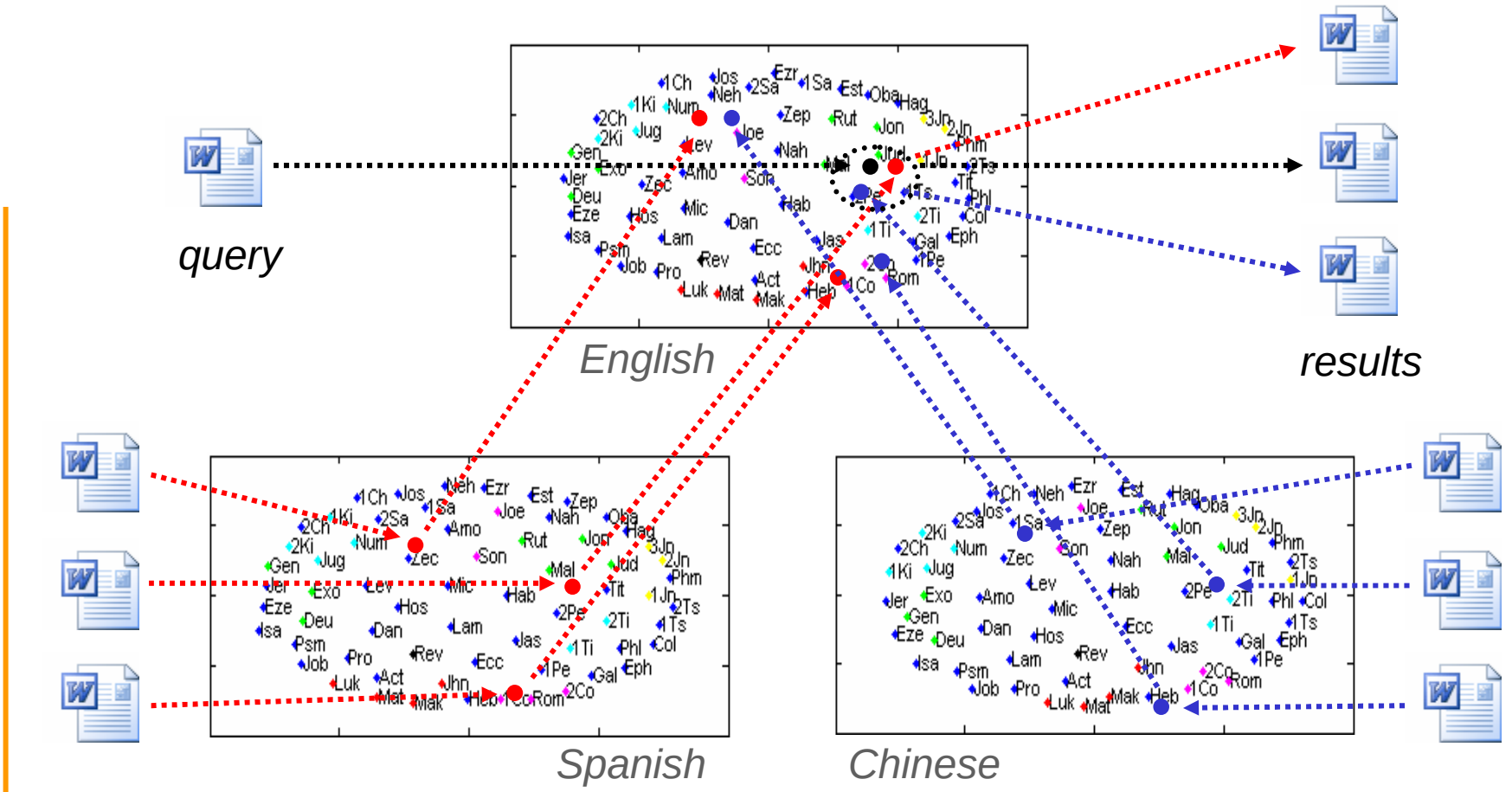
- The similarities among the maps are remarkable

- **Could we exploit these similarities for performing cross-language information retrieval tasks?**

Barcelona Media
Centre d'Innovació

The proposed technique

CLIR by using MDS projections



Barcelona Media
Centre d'Innovació

The proposed technique

- Start from a multilingual collection of “anchor documents” and construct the retrieval language map

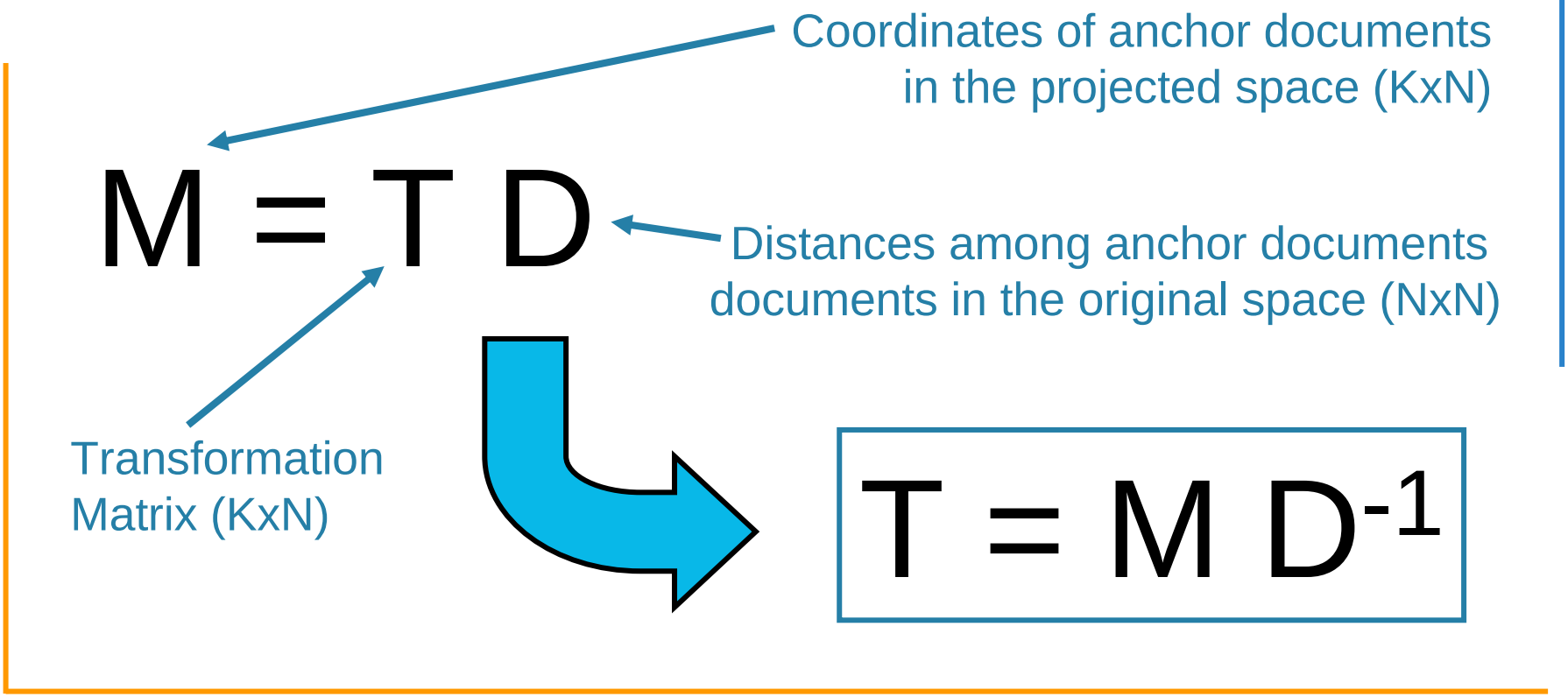
- Project new documents and queries from any source language into the retrieval language map:
  - Use linear projections (the one discussed here)
  - Use an optimization procedure (future work…)

- Retrieve documents over retrieval language map by using a distance metric

Barcelona Media
Centre d'Innovació

Projection matrix

A linear transformation from the original high dimensional space into the lower dimensionality map can be inferred from anchor documents



Two variants of T

Two different variants of the linear projection matrix T can be computed:

- A monolingual projection matrix:

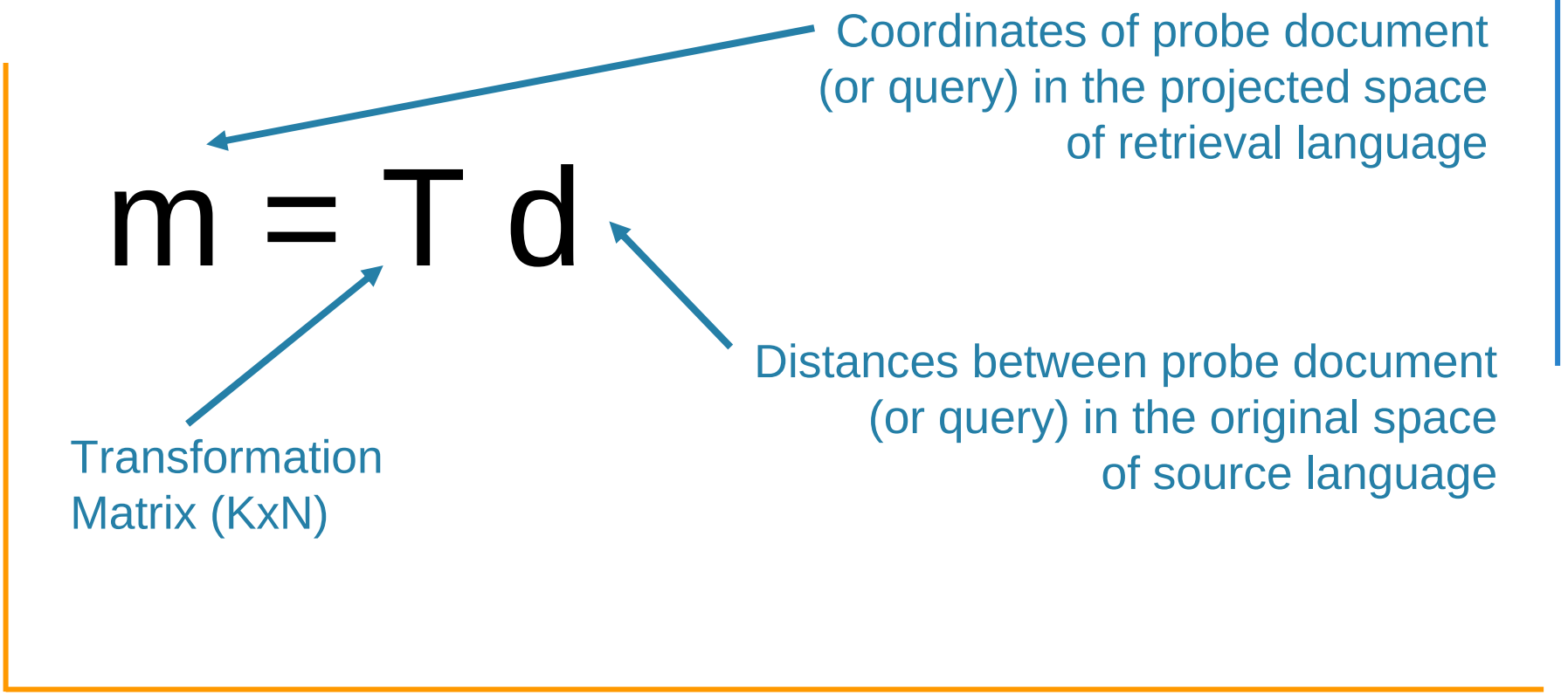
  M and **D** are computed on the retrieval language

- A cross-lingual projection matrix:

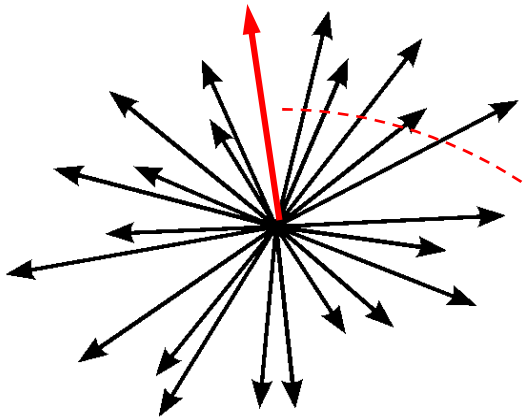
  M is computed on the retrieval language
  D is computed on the source language

Projecting new documents

A probe document or query can be placed into the retrieval map by using the transformation matrix



Projecting new documents



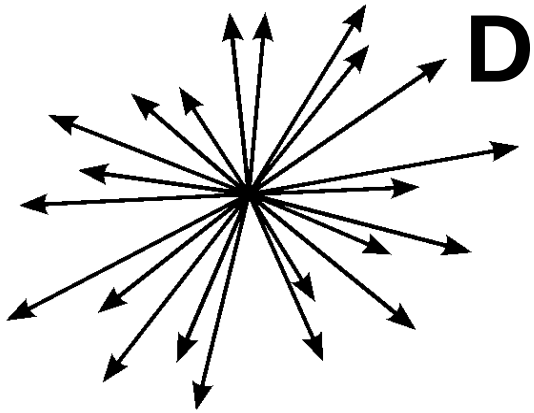
Source language

$$\mathbf{m = (MD^{-1})\ d}$$

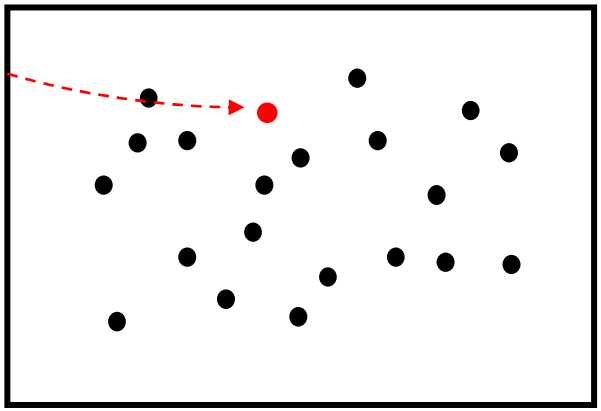
Monolingual projection matrix

M

D



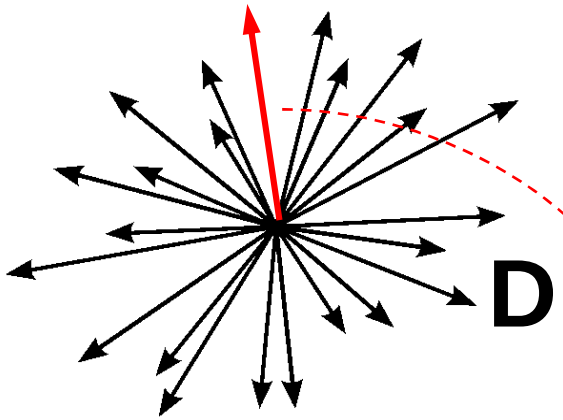
MDS



Retrieval language

Retrieval map

Projecting new documents



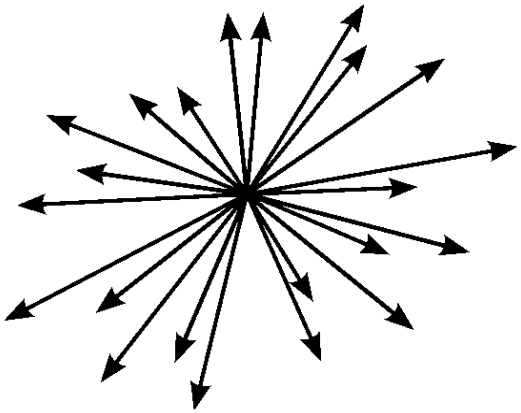
Source language

D

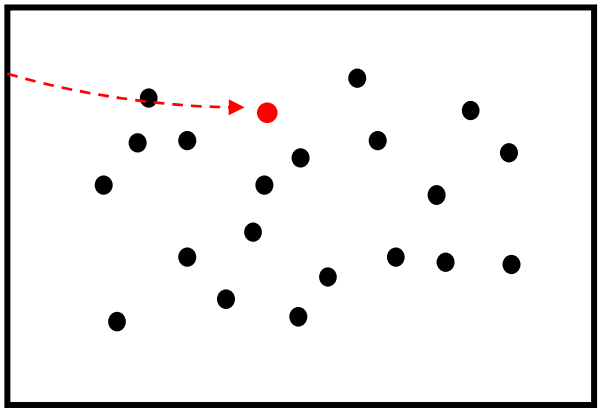
$$\mathbf{m = (MD^{-1})\ d}$$

Cross-lingual projection matrix

M



MDS



Retrieval language

Retrieval map

Barcelona Media
Centre d'Innovació

Latent Semantic Indexing

Latent Semantic Indexing for CLIR ***(Berry and Young, 1995)***

- In monolingual LSI, the term-document matrix is decomposed into a set of K orthogonal factors by means of Singular Value Decomposition (SVD)

- In cross-language LSI, a multilingual term-document matrix is constructed from a multilingual parallel collection and LSI is applied by considering multilingual "extended" representations of query and documents

Barcelona Media
Centre d'Innovació

Latent Semantic Indexing

Multilingual term-document matrix

$$X = \begin{bmatrix} Xa \\ Xb \end{bmatrix}$$

Xa: Term-document matrix in language A

Xb: Term-document matrix in language B

Singular Value Decomposition of X: $X = U \Sigma V^T$

X : MxN (M=Ma+Mb)
U : MxM (orthonormal "output" basis vector directions)
Σ : MxN (diagonal matrix containing P singular values)
V : NxN (orthonormal “input" basis vector directions)
P : min(M,N)

Retrieval is based on internal product of the form $\langle U^T d, U^T q \rangle$

$$d = \begin{bmatrix} da \\ 0 \end{bmatrix} \text{ ó } \begin{bmatrix} 0 \\ db \end{bmatrix} \qquad q = \begin{bmatrix} qa \\ 0 \end{bmatrix} \text{ ó } \begin{bmatrix} 0 \\ qb \end{bmatrix}$$

Barcelona Media
Centre d'Innovació

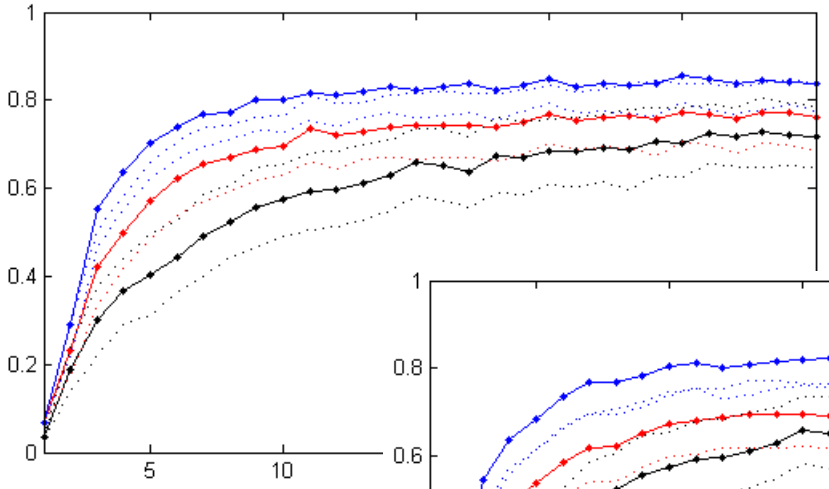
CLIR experiment 1

Retrieval task: retrieve a book by using the same book in a different language as a query

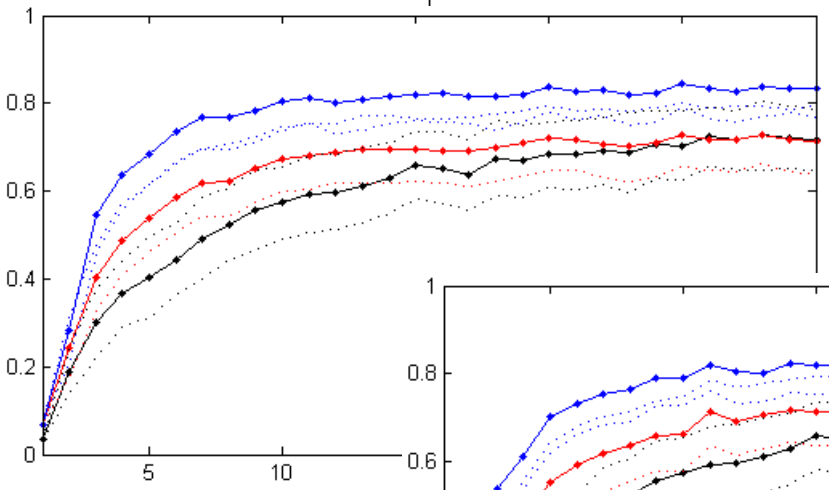
- Two variants of proposed method are compared with LSI
- Training set (anchor documents): 30 books
- Test set (documents to be retrieved): 30 books
- No overlap between training and test set
- Dimensionality of retrieval space is varied from 2 to 30
- 100 random selections for training and test sets at each run
- Average retrieval accuracy (top-1) is measured for each run
- Both English-Chinese retrieval tasks are considered

Barcelona Media
Centre d'Innovació

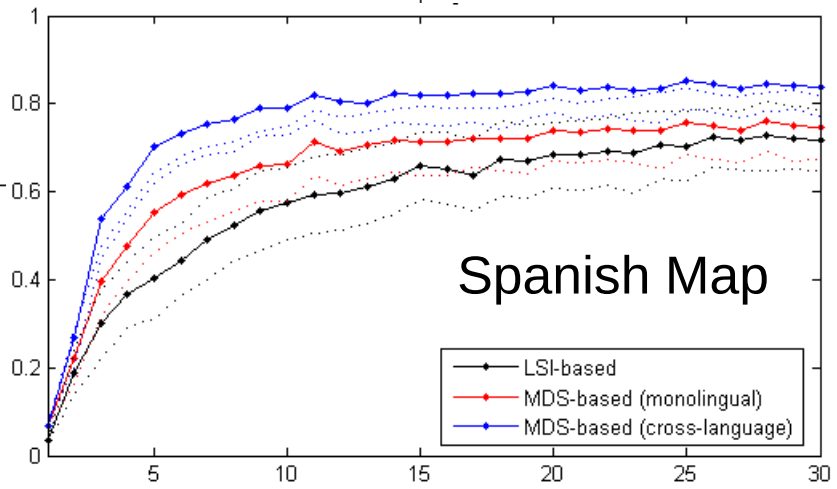
CLIR experiment 1



Retrieval carried out over Chinese Map



English Map

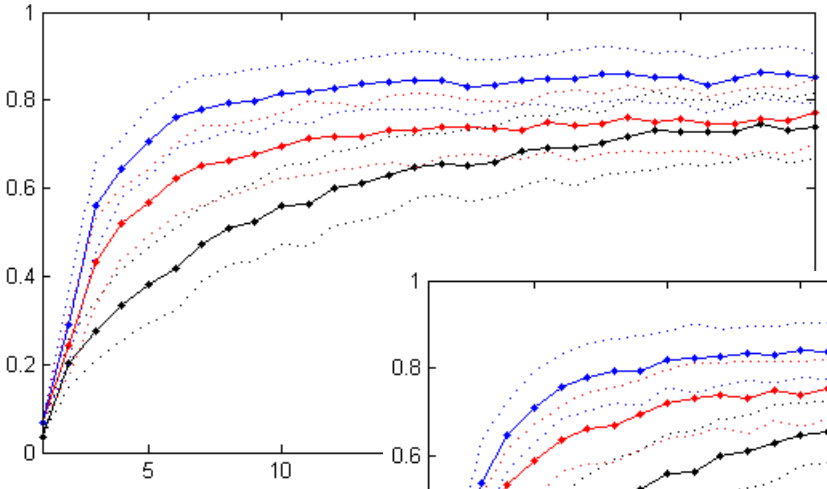


Spanish Map

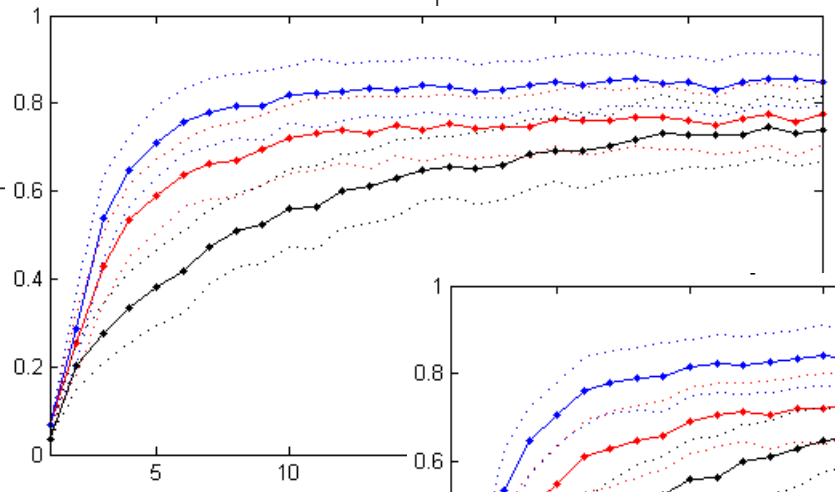
Chinese to English

Barcelona Media
Centre d'Innovació

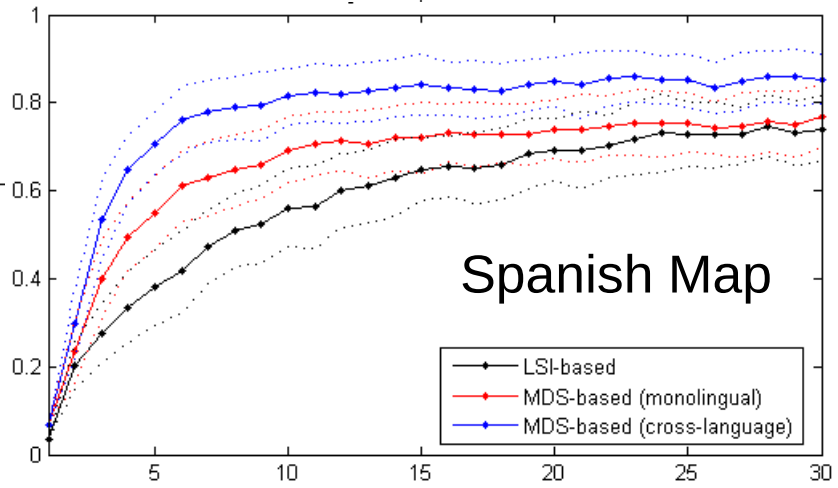
CLIR experiment 1



Retrieval carried out over Chinese Map



English Map



Spanish Map

English to Chinese

Barcelona Media
Centre d'Innovació

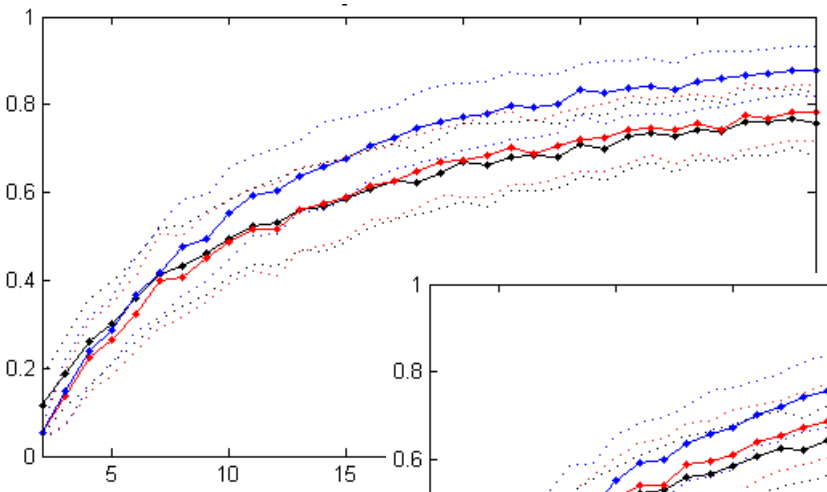
CLIR experiment 2

Retrieval task: retrieve a book by using the same book in a different language as a query

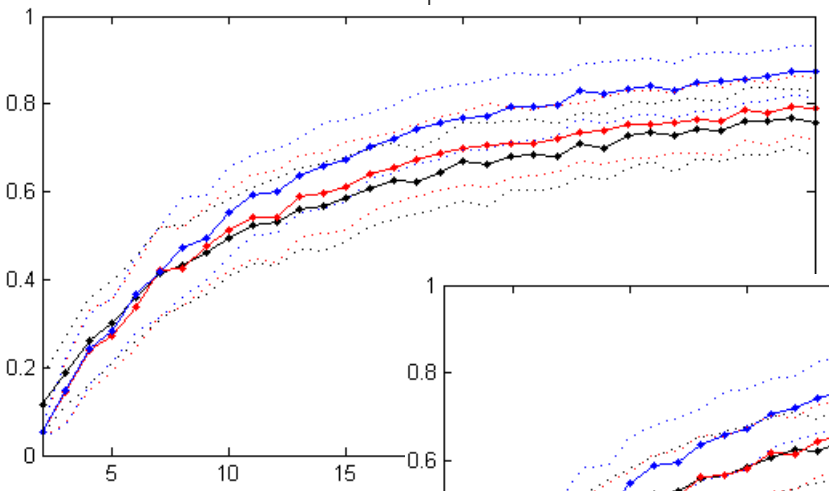
- Two variants of proposed method are compared with LSI
- Training set (anchor documents): vary from 2 to 35 books
- Test set (documents to be retrieved): 30 books
- No overlap between training and test set
- Dimensionality of retrieval space equal to training set size
- 100 random selections for training and test sets at each run
- Average retrieval accuracy (top-1) is measured for each run
- Only English to Chinese retrieval task is considered

Barcelona Media
Centre d'Innovació

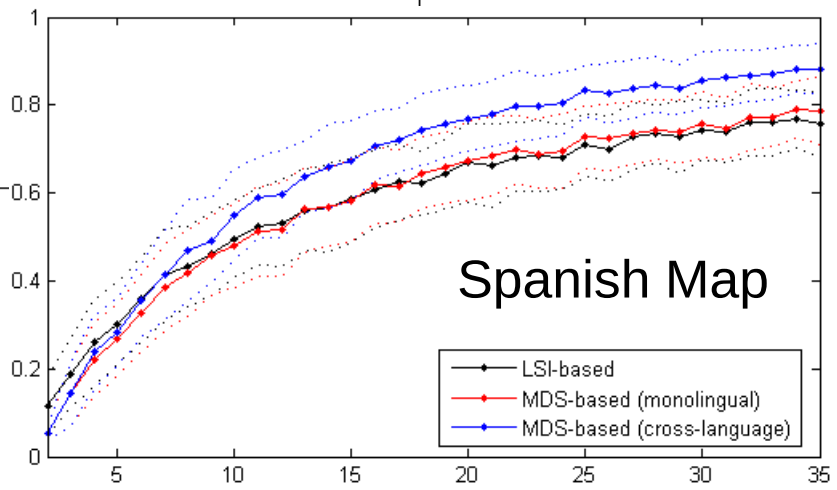
CLIR experiment 2



Retrieval carried out over Chinese Map



English Map



Spanish Map

English to Chinese

Barcelona Media
Centre d'Innovació

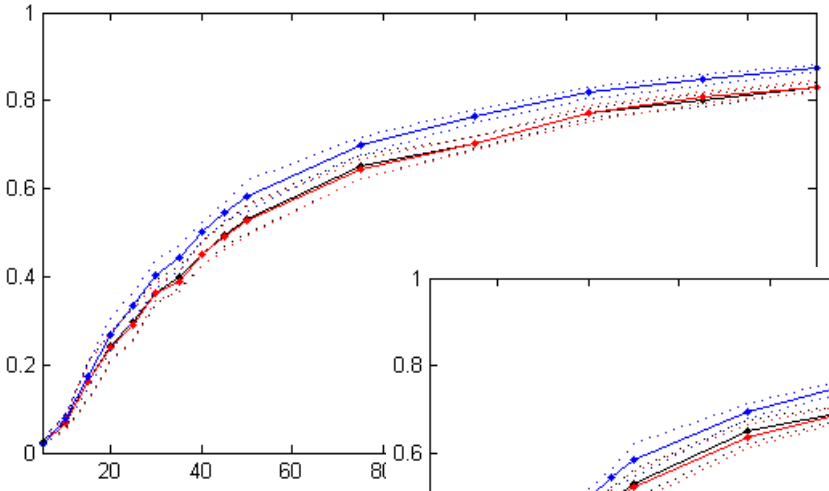
CLIR experiment 3

Retrieval task: retrieve a chapter by using the same chapter in a different language as a query

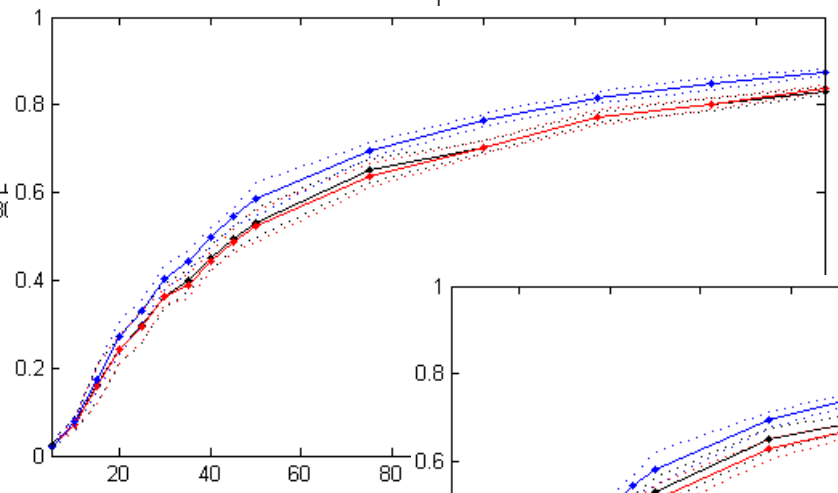
- Two variants of proposed method are compared with LSI
- Training set (anchor documents): vary from 5 to 175 chapters
- Test set (documents to be retrieved): 30 chapters
- No overlap between training and test set
- Dimensionality of retrieval space equal to training set size
- 100 random selections for training and test sets at each run
- Average retrieval accuracy (top-1) is measured for each run
- Only English to Chinese retrieval task is considered

Barcelona Media
Centre d'Innovació

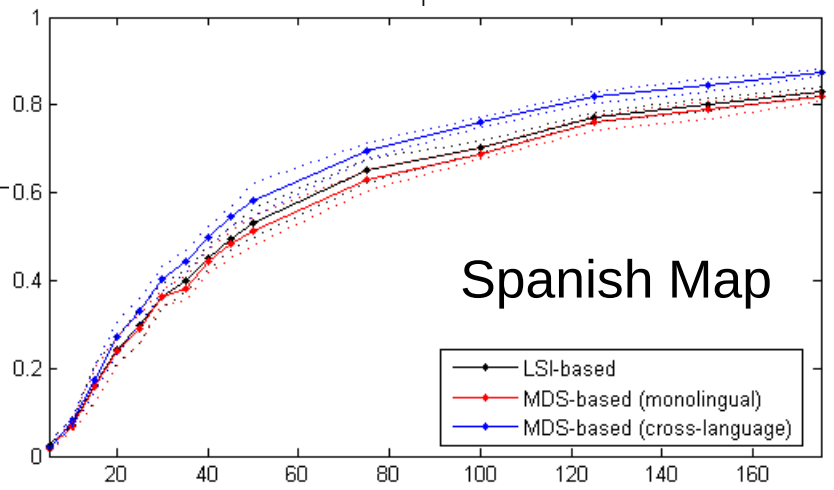
CLIR experiment 3



Retrieval carried out over Chinese Map



English Map



Spanish Map

English to Chinese

Barcelona Media
Centre d'Innovació

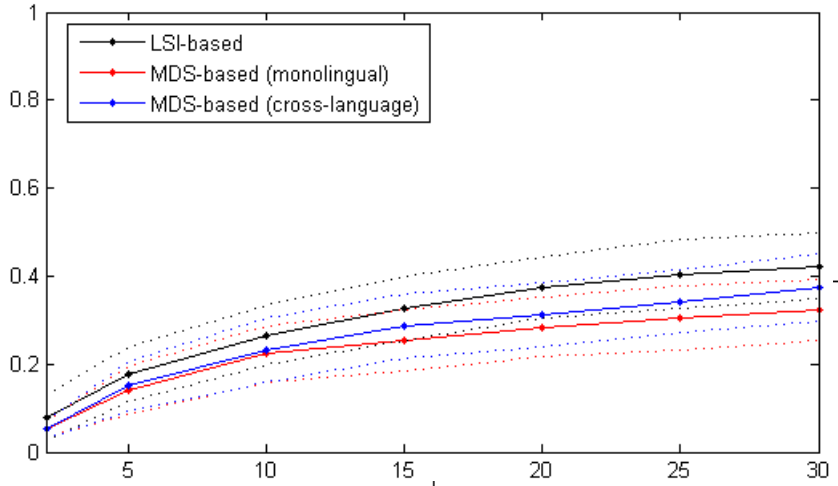
CLIR experiment 4

Retrieval task: retrieve a book by using any chapter of the same book in a different language as a query

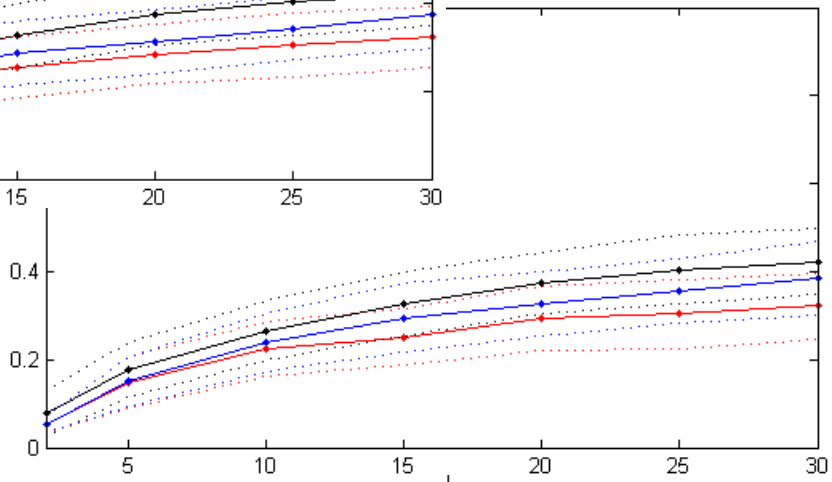
- Two variants of proposed method are compared with LSI
- Training set (anchor documents): vary from 2 to 30 books
- Test set (documents to be retrieved): 30 books
- No overlap between training and test set
- Dimensionality of retrieval space equal to training set size
- 100 random selections for training and test sets at each run
- Average retrieval accuracy (top-1) is measured for each run
- Only English to Chinese retrieval task is considered

Barcelona Media
Centre d'Innovació

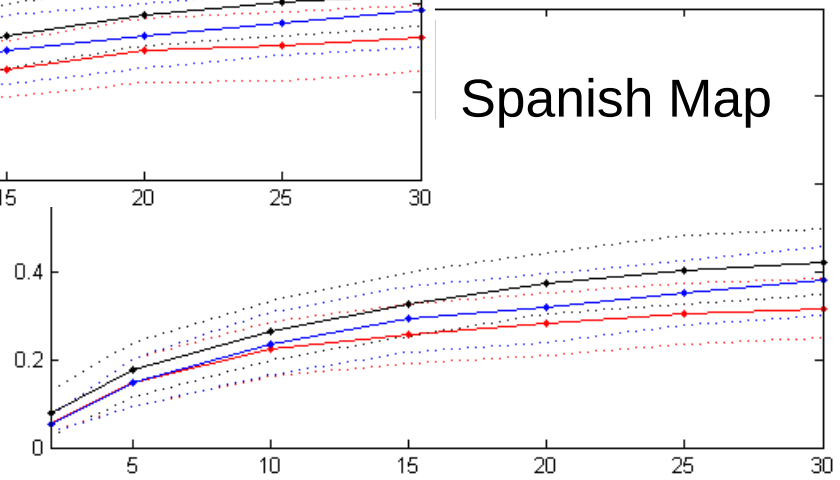
CLIR experiment 4



Retrieval carried out over Chinese Map



English Map



Spanish Map

English to Chinese

Barcelona Media
Centre d'Innovació

Conclusions

Some conclusions derived from these preliminary experiments:

- Preliminary results show that MDS projections can be exploited for CLIR tasks
- The cross-lingual projection matrix variant performs better that the monolingual projection matrix variant
- The technique performs significantly better than LSI when full text is used for retrieval,
- but it performs worse than LSI when partial text is used for retrieval…

Barcelona Media
Centre d'Innovació

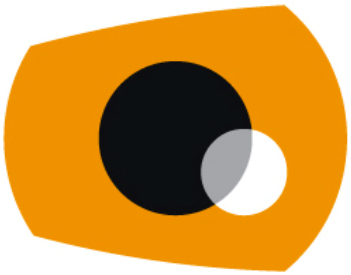
Future work

Future work in these research line involves:

- Understanding why the method performance deteriorates so much when partial text is used for retrieval

- Implementing an optimization procedure for probe document and query placement

- Explore alternatives for combining retrieval outputs when using different retrieval language maps



Centre
d'Innovació

22 Barcelona Media

QUESTIONS...