Barcelona Media
Centre d'Innovació

10th International Conference on Intelligent Text Processing and Computational Linguistics

Semantic Mapping for Related Term Identification

Rafael E. Banchs
Barcelona Media Innovation Centre, Spain

CICLing 2009, Mexico City, 3/3/09

Barcelona Media
Centre d'Innovació

Motivation

- Vector model space extensively used
- Latent semantic analysis:
  - more efficient representation (space reduction)
  - but still high dimensionality (400 – 600)
- Multidimensional Scaling:
  - non-linear projection method (comp. expensive)
  - preserves structural properties at lower dimensions
- **Could we combine advantages of both methods?**

Barcelona Media
Centre d'Innovació

Objectives

- To explore the advantages of the combined use of LSA and MDS
- To propose an experimental framework for text mining applications
- To illustrate the proposed methodology with a given application:
    - related term identification,

Latent Semantic Indexing

- Space reduction technique based on SVD (singular value decomposition)

$$D = U \Sigma V^T \Rightarrow U^T D = \Sigma V^T = Y$$

Projected Data (transformed space)

Original Data

$$Y = U^T D$$

Projected Data (reduced space)

$$Yc = Uc^T X$$

Optimal linear compression

$$U = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{m1} \\ u_{21} & u_{22} & \dots & u_{m2} \\ \vdots & \vdots & & \vdots \\ u_{m1} & u_{m2} & \dots & u_{mm} \end{pmatrix}$$

Uc

Barcelona Media
Centre d'Innovació

Multidimensional Scaling

A method for data visualization

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

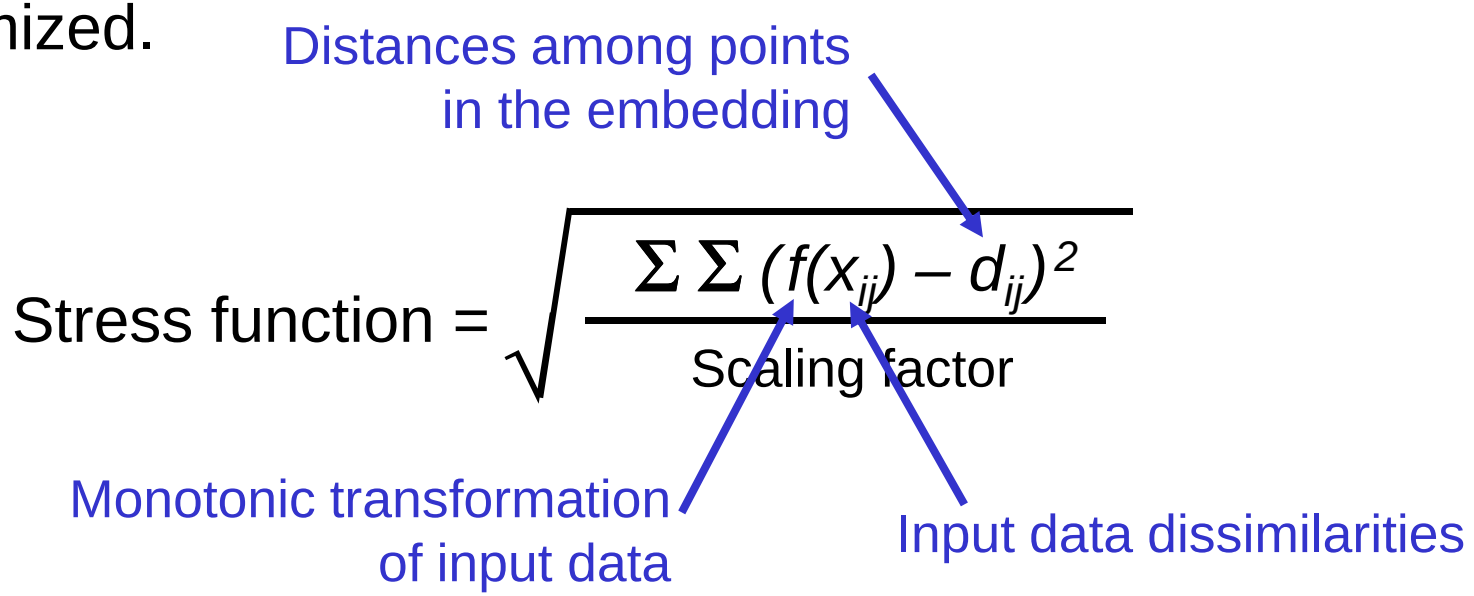
- Find a set of Euclidean coordinates to the objects in the group (i.e. an embedding)

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

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.



Barcelona Media
Centre d'Innovació

Proposed methodology

data collection

TF-IDF

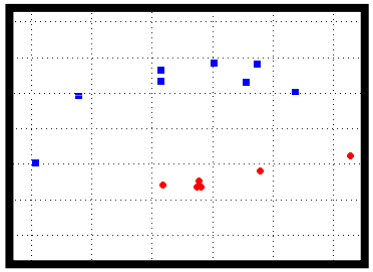
original-space representation

LSI

intermediate-space representation

MDS

low-dimensional semantic map

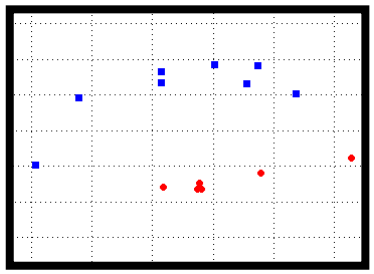


K-means

data clusters

MDS

low-dimensional semantic map



Illustrative examples

The proposed methodology is illustrated in the context of

- Related term identification
  - Individual term projections
    1.- effects of intermediate space dimensionality
    2.- final dimension selection
    3.- projection of new terms
  - Term cluster projections

Barcelona Media
Centre d'Innovació

Dataset description

Term definitions from a Spanish dictionary:

Collection	Terms	Definitions	Aver. Length
Verbs	4,800	12,414	6.05 words
Adjectives	5,390	8,596	6.05 words
Nouns	20,592	38,689	9.56 words
Others	5,273	9,835	8.01 words
Complete	36,055	69,534	8.32 words

- Each definition treated as a document
- Original space dimensionality 12,913 reduced to 7,198 after eliminating singletons
- Standard TF-IDF weighting and normalization

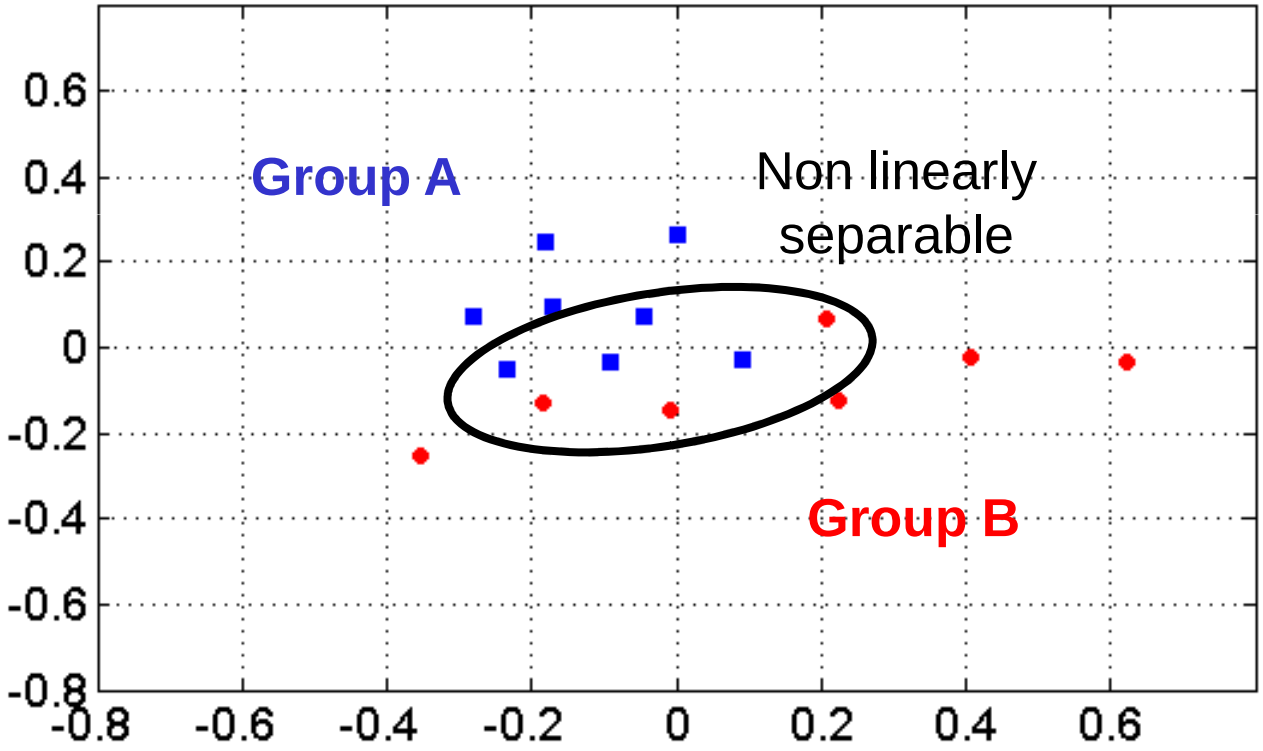
Projection of terms

Consider two groups of semantically related verbs:

Group A	**Group B**
Ayudar (to help)	Agredir (to threaten)
Compartir (to share)	Destruir (to destroy)
Beneficiar (to benefit)	Aniquilar (to eliminate)
Colaborar (to collaborate)	Atacar (to attack)
Salvar (to save)	Arruinar (to ruin)
Apoyar (to support)	Matar (to kill)
Cooperar (to cooperate)	Perjudicar (to perjudice)
Favorecer (to favour)	–

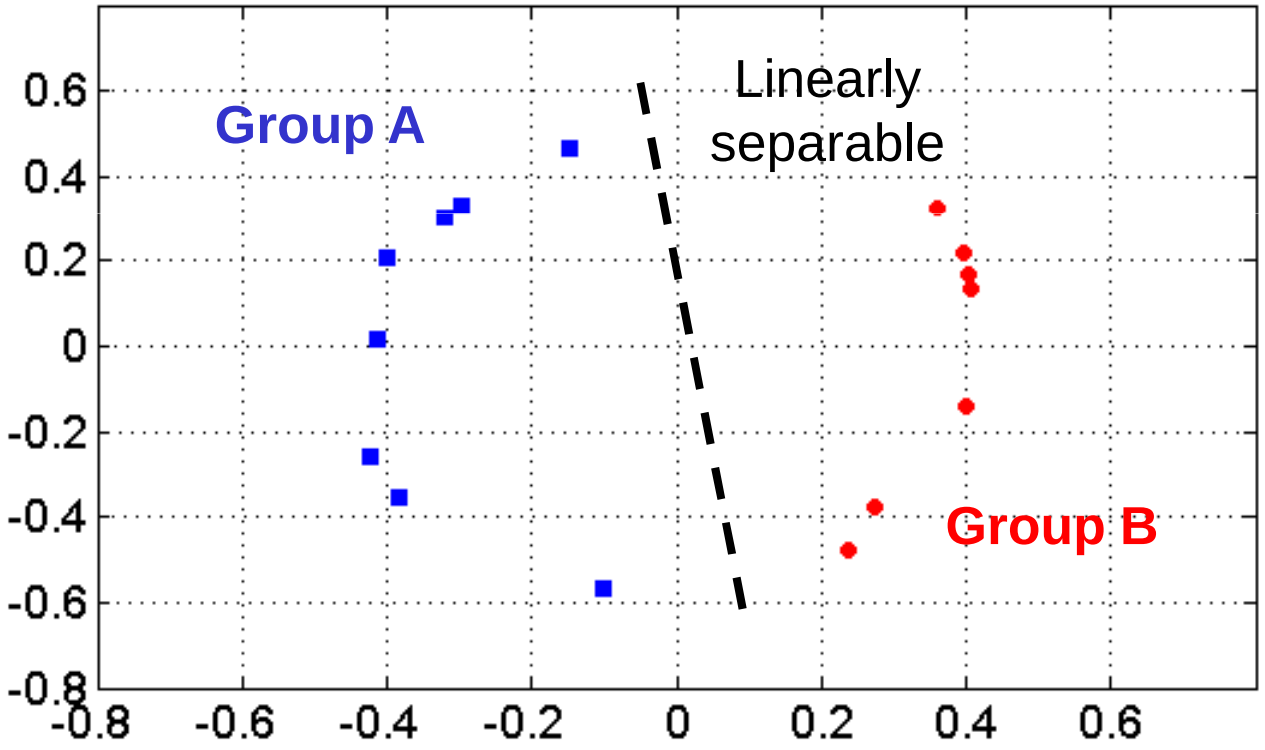
Intermediate dimension

Intermediate dimension = original dimension (NO LSI)



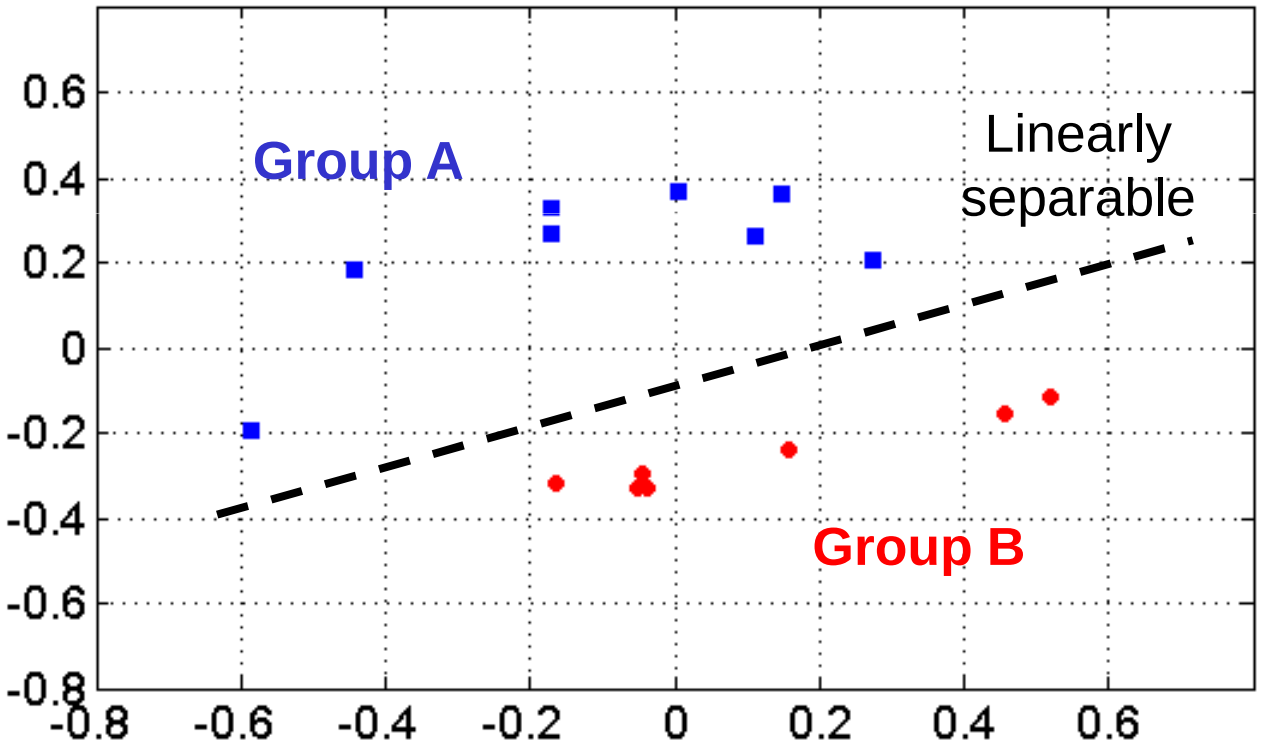
Intermediate dimension

Intermediate dimension = 800



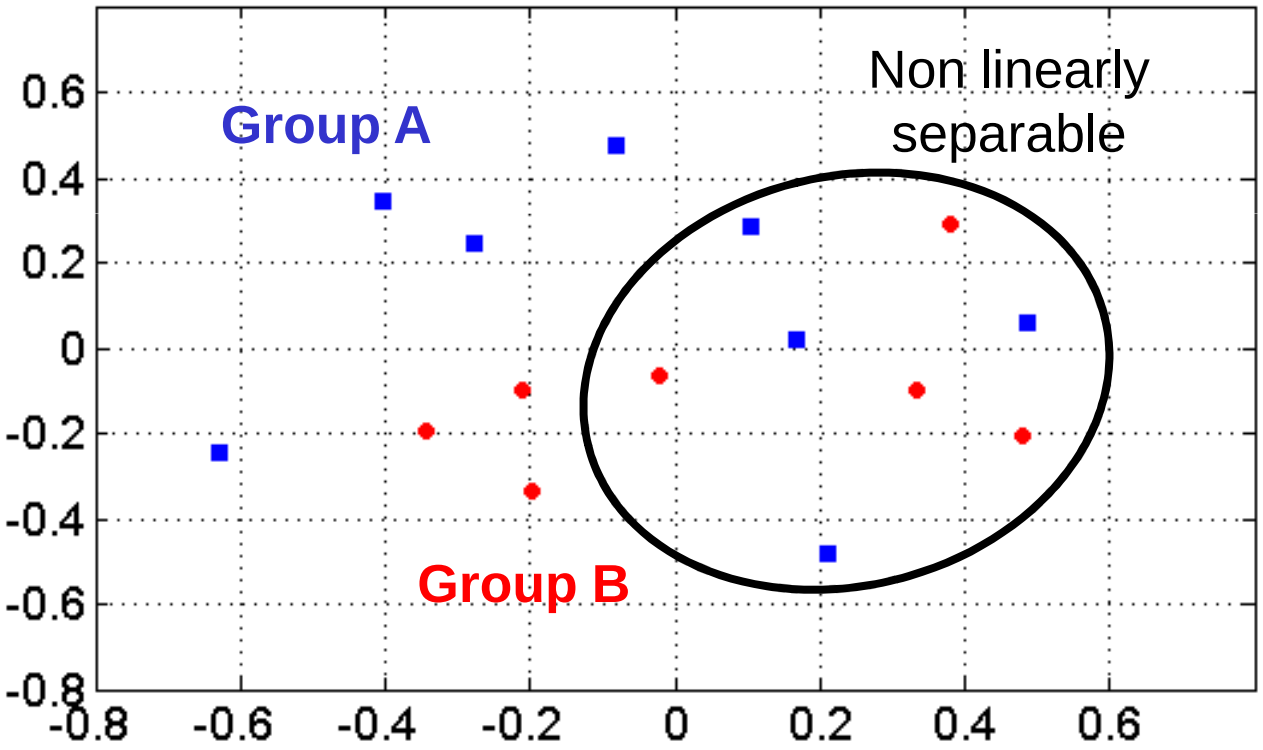
Intermediate dimension

Intermediate dimension = 400



Intermediate dimension

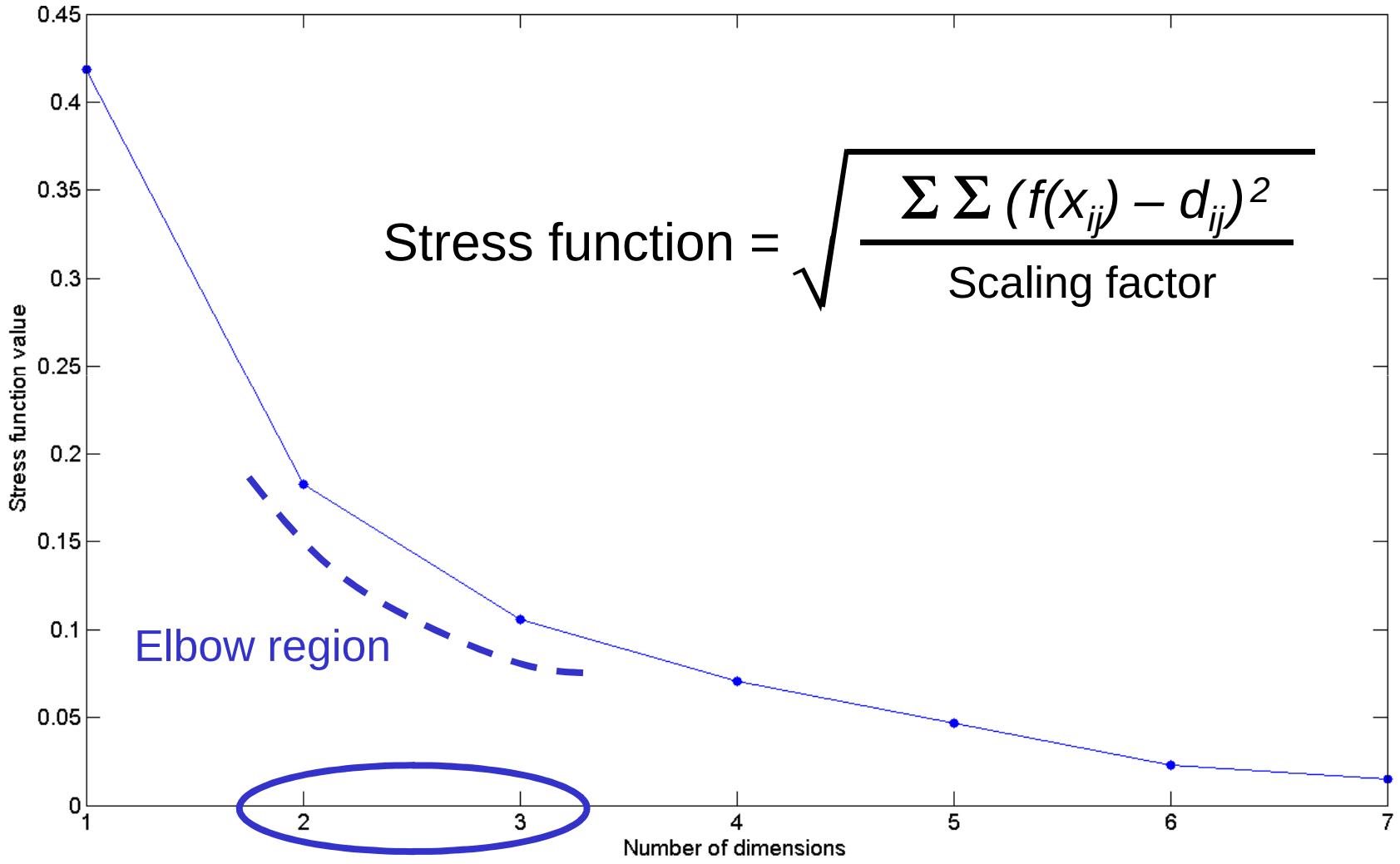
Intermediate dimension = 100



Barcelona Media
Centre d'Innovació

Final dimension selection

$$\text{Stress function} = \sqrt{\frac{\sum \sum (f(x_{ij}) - d_{ij})^2}{\text{Scaling factor}}}$$



Projecting new terms

Problem: MDS projections are data dependent, i.e. including a new element will generate a different map

Possible solution: to implement a linear mapping between the new map and the original map

Pseudo inverse operator

Coordinates of the new term in the new map

$$I_o = [\, M_o \, pinv(M_n) \,]\, I_n$$

Coordinates of original terms in the original map

Coordinates of original terms in the new map

Projecting new terms

Consider new terms semantically related to the original two groups:

Group A	**Group B**
Solidarizar (to solidarize)	Vencer (to defeat)
Apadrinar (to uphold)	Dominar (to dominate)
Proteger (to protect)	Someter (to enslave)
Defender (to defend)	Destrozar (to destroy)

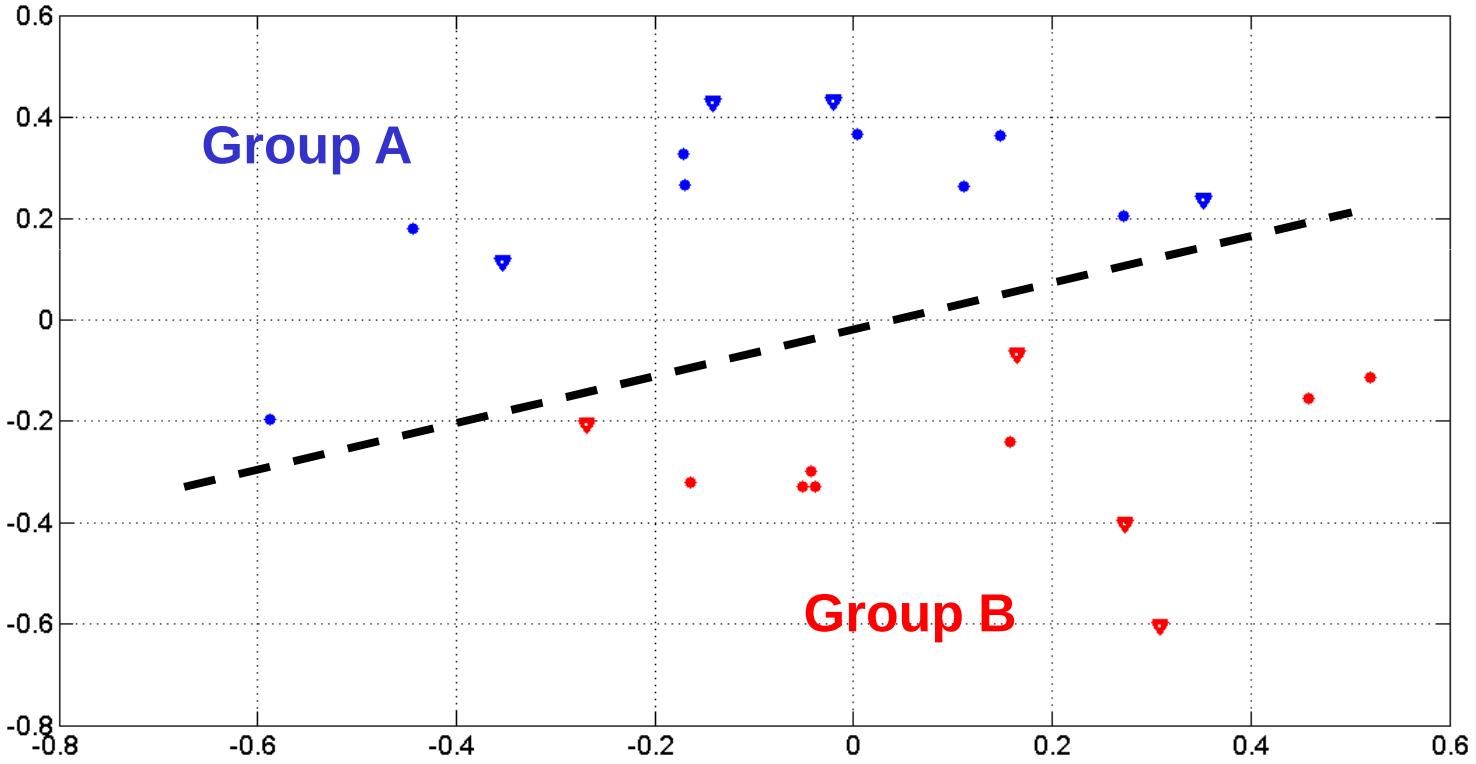
- Will be put on the original map by using the proposed linear mapping technique

Barcelona Media
Centre d'Innovació

Projecting new terms

Intermediate dimension = 400

▼ New terms

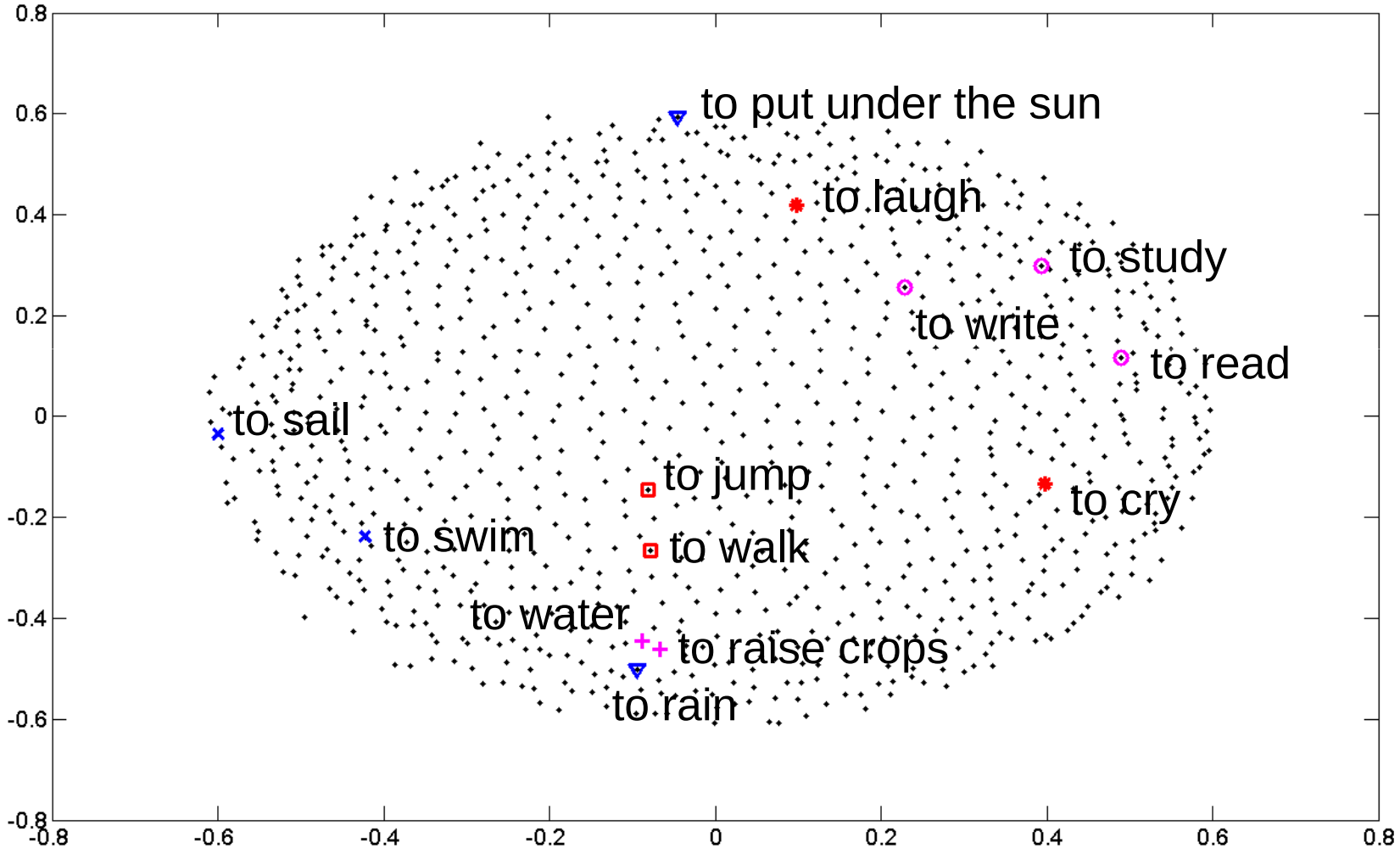


Mapping of clusters

- The totality of the 12,414 entries for verbs were considered
- An 800-dimensional intermediate space representation was generated
- k-means was applied to group the 12,414 entries into 1,000 clusters
    - *Minimum size 2 and maximum size 36*
    - *Mean size 12.4 and variance 4.7*
- Finally, MDS was applied to generate a map

Barcelona Media
Centre d'Innovació

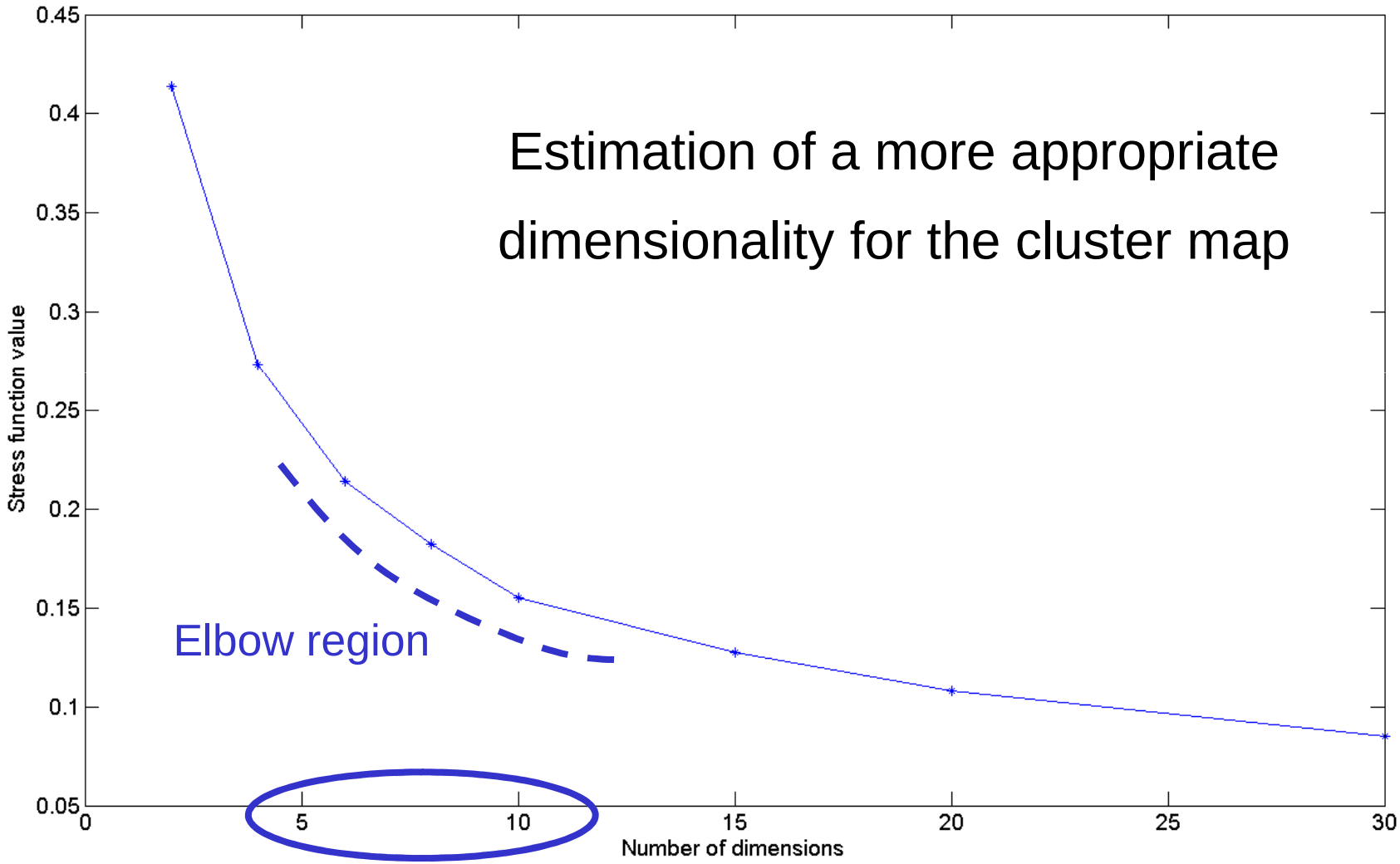
Mapping clusters



Barcelona Media
Centre d'Innovació

Mapping clusters

Estimation of a more appropriate dimensionality for the cluster map



Conclusions and future work

- Combining LSA with a non linear projection method has proven to allow for further reducing space dimensionality while preserving structural characterization
- Some positive preliminary results have been already obtained for related term identification
- We intend to explore about the applicability and usefulness of the proposed method in other applications, specifically: document classification, topic detection and tracking, and opinion mining…

Barcelona Media
Centre d'Innovació

10th International Conference on Intelligent Text Processing and Computational Linguistics

Semantic Mapping for Related Term Identification

Rafael E. Banchs
Barcelona Media Innovation Centre, Spain

CICLing 2009, Mexico City, 3/3/09