Semantically rich spaces for document clustering

Roberto Basili   Paolo Marocco   Daniele Milizia

DISP
University of Rome Tor Vergata, Rome, Italy
{basili,marocco,milizia}@info.uniroma2.it

Text-based IR Workshop

DEXA 2008, Torino, September 1st 2008

Outline

Outline

Outline

Outline

Document Data and Language Learning

- Electronic Documents embody massive information about language **in use**

Document Data and Language Learning

- Electronic Documents embody massive information about language **in use**
- This makes automatic extaction interesting for acquiring/adapting large scale components of lexical knowledge bases

Document Data and Language Learning

- Electronic Documents embody massive information about language **in use**
- This makes automatic extaction interesting for acquiring/adapting large scale components of lexical knowledge bases
- *Data sparseness* is amplified by *language variability*

Document Data and Language Learning

- Electronic Documents embody massive information about language **in use**
- This makes automatic extaction interesting for acquiring/adapting large scale components of lexical knowledge bases
- *Data sparseness* is amplified by *language variability*
- *Uncertainty* is amplified by *language ambiguity*

Lexical Learning and Vector Spaces

- Semantic Information is needed in several lexical tasks (e.g. Question Answering)
- Vectors are usually representing words, word senses, patterns, or even predicates (such as in Framenet)
- Weights characterize topical, syntagmatic or paradigmatic features

Lexical Learning and Vector Spaces

- Semantic Information is needed in several lexical tasks (e.g. Question Answering)
- Vectors are usually representing words, word senses, patterns, or even predicates (such as in Framenet)
- Weights characterize topical, syntagmatic or paradigmatic features

Challanges

- Representation: which features are best suited for the target linguistic elements

Lexical Learning and Vector Spaces

- Semantic Information is needed in several lexical tasks (e.g. Question Answering)
- Vectors are usually representing words, word senses, patterns, or even predicates (such as in Framenet)
- Weights characterize topical, syntagmatic or paradigmatic features

Challanges

- Representation: which features are best suited for the target linguistic elements
- Induction: which *similarity* function is to be modeled in the different spaces
- Inference: which operators define suitable *compositional deductions*

Local and global infomation in DR methods

Dimensionality Reduction methods explore the data distribution properties for minimizing the number of features needed for reaching good levels of accuracy.

- Work in ML explores the impact of DR methods based on function metrics evoked by the data distribution themselves

Local and global infomation in DR methods

Dimensionality Reduction methods explore the data distribution properties for minimizing the number of features needed for reaching good levels of accuracy.

- Work in ML explores the impact of DR methods based on function metrics evoked by the data distribution themselves
  - Linear Discriminant Analysis
  - Spectral Clustering
  - LPP
- This helps in minimizing the impact of data sparseness, improving complexity as well as keeping accuracy at reasonable levels

Local and global infomation in DR methods

Dimensionality Reduction methods explore the data distribution properties for minimizing the number of features needed for reaching good levels of accuracy.

- Work in ML explores the impact of DR methods based on function metrics evoked by the data distribution themselves
  - Linear Discriminant Analysis
  - Spectral Clustering
  - LPP
- This helps in minimizing the impact of data sparseness, improving complexity as well as keeping accuracy at reasonable levels
- These formulations give rise to valid kernels highly interesting for CoNLL

Local and global infomation in DR methods

Objectives

- To compare and validate data-driven metrics on realistic tasks
- To validate the linguistic information provided by the corresponding spaces
- To determine kernels relevant for CoNLL research

Semantic spaces

Semantic Spaces: a definition

A Semantic Space for a set of N targets is 4-tuple $< B, A, S, V >$ where

- B is the set of basic features (e.g. words co-occurring with the targets)

Semantic spaces

Semantic Spaces: a definition

A Semantic Space for a set of N targets is 4-tuple $< B, A, S, V >$ where

- B is the set of basic features (e.g. words co-occurring with the targets)
- A is a lexical association function that weights the correlations between $b \in B$ and the targets

Semantic spaces

Semantic Spaces: a definition

A Semantic Space for a set of N targets is 4-tuple $< B, A, S, V >$ where

- B is the set of basic features (e.g. words co-occurring with the targets)
- A is a lexical association function that weights the correlations between $b \in B$ and the targets
- S is a similarity function between targets (i.e. in $\Re^{|B|} \times \Re^{|B|}$)

Semantic spaces

Semantic Spaces: a definition

A Semantic Space for a set of N targets is 4-tuple $< B, A, S, V >$ where

- B is the set of basic features (e.g. words co-occurring with the targets)
- A is a lexical association function that weights the correlations between $b \in B$ and the targets
- S is a similarity function between targets (i.e. in $\Re^{|B|} \times \Re^{|B|}$)
- V is a linear transformation over the original $N \times |B|$ matrix

Semantic spaces

Semantic Spaces: a definition

A Semantic Space for a set of N targets is 4-tuple $< B, A, S, V >$ where

- B is the set of basic features (e.g. words co-occurring with the targets)
- A is a lexical association function that weights the correlations between $b \in B$ and the targets
- S is a similarity function between targets (i.e. in $\Re^{|B|} \times \Re^{|B|}$)
- V is a linear transformation over the original $N \times |B|$ matrix

Examples

- In IR systems, targets are documents, B is the term vocabulary, A is the $tf \cdot idf$ score. The S function is usually the cosine similarity, i.e. $sim(\vec{t_1}, \vec{t_2}) = \frac{\sum_i t_{1i} \cdot t_{2i}}{||\vec{t_1}|| \cdot ||\vec{t_2}||}$

Semantic spaces

Semantic Spaces: a definition

A Semantic Space for a set of N targets is 4-tuple $< B, A, S, V >$ where

- B is the set of basic features (e.g. words co-occurring with the targets)
- A is a lexical association function that weights the correlations between $b \in B$ and the targets
- S is a similarity function between targets (i.e. in $\Re^{|B|} \times \Re^{|B|}$)
- V is a linear transformation over the original $N \times |B|$ matrix

Examples

- In IR systems, targets are documents, B is the term vocabulary, A is the $tf \cdot idf$ score. The S function is usually the cosine similarity, i.e. $sim(\vec{t_1}, \vec{t_2}) = \frac{\sum_i t_{1i} \cdot t_{2i}}{||\vec{t_1}|| \cdot ||\vec{t_2}||}$
- In Latent Semantic Analysis (Berry et al. 94) targets can be documents or words, and the transformation V is SVD

Latent Semantic Spaces

LSA and Lexical semantics

In LSA approaches, SVD is applied to source co-occurrence matrices in order to

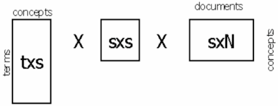
- Reduce the original dimensionality
- Capture *topical similarity* latent in the original documents, i.e. second order relations among targets (words)

M =

	d_1	d_2	d_3	d_4	d_5	d_6
shuttle	1	0	1	0	0	0
astronaut	0	1	0	0	0	0
moon	1	1	0	0	0	0
car	1	0	0	1	1	0
truck	0	0	0	1	0	1

$$\mathbf{M} = \mathbf{K}_{t\times s}\mathbf{S}_{s\times s}\mathbf{D}^T_{s\times N}$$

=



K =

	dim_1	dim_2	dim_3	dim_4	dim_5
shuttle	-0.44	-0.30	0.57	0.58	0.25
astronaut	-0.13	-0.33	-0.59	0.00	0.73
moon	-0.48	-0.51	-0.37	0.00	-0.61
car	-0.70	0.35	0.15	-0.58	0.16
truck	-0.26	0.65	-0.41	0.58	-0.09

S =

2.16	0.00	0.00	0.00	0.00
0.00	1.59	0.00	0.00	0.00
0.00	0.00	1.28	0.00	0.00
0.00	0.00	0.00	1.00	0.00
0.00	0.00	0.00	0.00	0.39

Latent Semantic Spaces

LSA and Lexical semantics

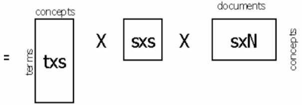
In LSA approaches, SVD is applied to source co-occurrence matrices in order to

- Reduce the original dimensionality
- Capture *topical similarity* latent in the original documents, i.e. second order relations among targets (words)

M =

	d_1	d_2	d_3	d_4	d_5	d_6
shuttle	1	0	1	0	0	0
astronaut	0	1	0	0	0	0
moon	1	1	0	0	0	0
car	1	0	0	1	1	0
truck	0	0	0	1	0	1

$$\mathbf{M} = \mathbf{K}_{t\times s}\mathbf{S}_{s\times s}\mathbf{D}^T_{s\times N}$$



K =

	dim_1	dim_2	dim_3	dim_4	dim_5
shuttle	-0.44	-0.30	0.57	0.58	0.25
astronaut	-0.13	-0.33	-0.59	0.00	0.73
moon	-0.48	-0.51	-0.37	0.00	-0.61
car	-0.70	0.35	0.15	-0.58	0.16
truck	-0.26	0.65	-0.41	0.58	-0.09

S =

2.16	0.00	0.00	0.00	0.00
0.00	1.59	0.00	0.00	0.00
0.00	0.00	1.28	0.00	0.00
0.00	0.00	0.00	1.00	0.00
0.00	0.00	0.00	0.00	0.39

Latent Semantic Spaces

LSA and Lexical semantics

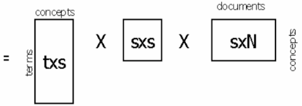
In LSA approaches, SVD is applied to source co-occurrence matrices in order to

- Reduce the original dimensionality
- Capture *topical similarity* latent in the original documents, i.e. second order relations among targets (words)

M =

	d_1	d_2	d_3	d_4	d_5	d_6
shuttle	1	0	1	0	0	0
astronaut	0	1	0	0	0	0
moon	1	1	0	0	0	0
car	1	0	0	1	1	0
truck	0	0	0	1	0	1

$$\mathbf{M} = \mathbf{K}_{t \times s} \mathbf{S}_{s \times s} \mathbf{D}^T_{s \times N}$$



K =

	dim_1	dim_2	dim_3	dim_4	dim_5
shuttle	-0.44	-0.30	0.57	0.58	0.25
astronaut	-0.13	-0.33	-0.59	0.00	0.73
moon	-0.48	-0.51	-0.37	0.00	-0.61
car	-0.70	0.35	0.15	-0.58	0.16
truck	-0.26	0.65	-0.41	0.58	-0.09

S =

2.16	0.00	0.00	0.00	0.00
0.00	1.59	0.00	0.00	0.00
0.00	0.00	1.28	0.00	0.00
0.00	0.00	0.00	1.00	0.00
0.00	0.00	0.00	0.00	0.39

Latent Semantic Spaces

LSA and Lexical semantics

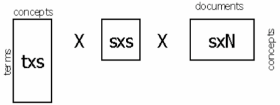
In LSA approaches, SVD is applied to source co-occurrence matrices in order to

- Reduce the original dimensionality
- Capture *topical similarity* latent in the original documents, i.e. second order relations among targets (words)

M =

	d_1	d_2	d_3	d_4	d_5	d_6
shuttle	1	0	1	0	0	0
astronaut	0	1	0	0	0	0
moon	1	1	0	0	0	0
car	1	0	0	1	1	0
truck	0	0	0	1	0	1

$$\mathbf{M} = \mathbf{K}_{t\times s}\mathbf{S}_{s\times s}\mathbf{D}^T_{s\times N}$$



K =

	dim_1	dim_2	dim_3	dim_4	dim_5
shuttle	-0.44	-0.30	0.57	0.58	0.25
astronaut	-0.13	-0.33	-0.59	0.00	0.73
moon	-0.48	-0.51	-0.37	0.00	-0.61
car	-0.70	0.35	0.15	-0.58	0.16
truck	-0.26	0.65	-0.41	0.58	-0.09

S =

2.16	0.00	0.00	0.00	0.00
0.00	1.59	0.00	0.00	0.00
0.00	0.00	1.28	0.00	0.00
0.00	0.00	0.00	1.00	0.00
0.00	0.00	0.00	0.00	0.39

Latent Semantic Spaces

LSA and Lexical semantics

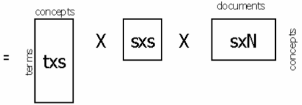
In LSA approaches, SVD is applied to source co-occurrence matrices in order to

- Reduce the original dimensionality
- Capture *topical similarity* latent in the original documents, i.e. second order relations among targets (words)

M

	d_1	d_2	d_3	d_4	d_5	d_6
shuttle	1	0	1	0	0	0
astronaut	0	1	0	0	0	0
moon	1	1	0	0	0	0
car	1	0	0	1	1	0
truck	0	0	0	1	0	1

$$\mathbf{M} = \mathbf{K}_{t\times s}\mathbf{S}_{s\times s}\mathbf{D}^T_{s\times N}$$



K =

	dim_1	dim_2	dim_3	dim_4	dim_5
shuttle	-0.44	-0.30	0.57	0.58	0.25
astronaut	-0.13	-0.33	-0.59	0.00	0.73
moon	-0.48	-0.51	-0.37	0.00	-0.61
car	-0.70	0.35	0.15	-0.58	0.16
truck	-0.26	0.65	-0.41	0.58	-0.09

S =

2.16	0.00	0.00	0.00	0.00
0.00	1.59	0.00	0.00	0.00
0.00	0.00	1.28	0.00	0.00
0.00	0.00	0.00	1.00	0.00
0.00	0.00	0.00	0.00	0.39

LSA: semantic interpretation

LSA and PCA

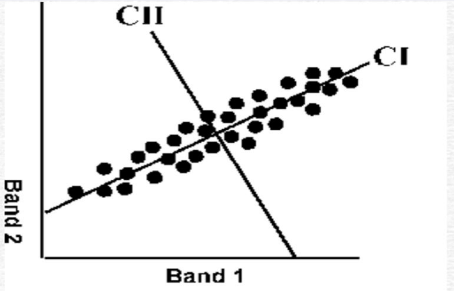


Figure 1 : The Principal Components Transform
[after Lillesand and Kiefer, 1994, 573]

- SVD let the principal components of the distribution emerge (max covariance)
- Principal components are linear combinations of the original dimensions, i.e. pseudo concepts, as captured in the <u>entire</u> space

LPP as a data-driven metrics

General Idea

- Determine the *best* linear transformation **A** that preserves the *local* properties of the space, without making global assumptions (as in LSA)
- An adjecency graph **G** is adopted, based on internal metrics (i.e. the space inner product) or external ones (e.g. dictionaries)

LPP as a data-driven metrics

General Idea

- Determine the *best* linear transformation **A** that preserves the *local* properties of the space, without making global assumptions (as in LSA)
- An adjecency graph **G** is adopted, based on internal metrics (i.e. the space inner product) or external ones (e.g. dictionaries)

Formally:

- $\arg\min_{\mathbf{a}} \sum_{ij} (\mathbf{a}^T x_i - \mathbf{a}^T x_j)^2 W_{ij}$

LPP as a data-driven metrics

General Idea

- Determine the *best* linear transformation **A** that preserves the *local* properties of the space, without making global assumptions (as in LSA)
- An adjecency graph **G** is adopted, based on internal metrics (i.e. the space inner product) or external ones (e.g. dictionaries)

Formally:

- $\arg\min_{\mathbf{a}} \sum_{ij} (\mathbf{a}^T x_i - \mathbf{a}^T x_j)^2 W_{ij}$
- $D_{ii} = \sum_j W_{ij}$ and $L = D - W$ (Laplacian matrix)

LPP as a data-driven metrics

General Idea

- Determine the *best* linear transformation **A** that preserves the *local* properties of the space, without making global assumptions (as in LSA)
- An adjecency graph **G** is adopted, based on internal metrics (i.e. the space inner product) or external ones (e.g. dictionaries)

Formally:

- $\arg\min_{\mathbf{a}} \sum_{ij} (\mathbf{a}^T x_i - \mathbf{a}^T x_j)^2 W_{ij}$
- $D_{ii} = \sum_j W_{ij}$ and $L = D - W$ (Laplacian matrix)
- Solve the eingenvector problem: $XLX^T\mathbf{a} = \lambda XDX^T\mathbf{a}$

LPP as a data-driven metrics

General Idea

- Determine the *best* linear transformation **A** that preserves the *local* properties of the space, without making global assumptions (as in LSA)
- An adjecency graph **G** is adopted, based on internal metrics (i.e. the space inner product) or external ones (e.g. dictionaries)

Formally:

- $\arg\min_{\mathbf{a}} \sum_{ij} (\mathbf{a}^T x_i - \mathbf{a}^T x_j)^2 W_{ij}$
- $D_{ii} = \sum_j W_{ij}$ and $L = D - W$ (Laplacian matrix)
- Solve the eingenvector problem: $XLX^T \mathbf{a} = \lambda XDX^T \mathbf{a}$
- Final projection into $\Re^k$: $(Y)_{k \times k} = A^T X$

The Adjacency Graph, G

Given two vectors x_i and x_j, **G** defines weights w_{ij}, as:

- *cosine* **graph**: $w_{ij} = max\{0, \frac{cos(x_i,x_j)-\tau}{|cos(x_i,x_j)-\tau|} \cdot cos(x_i, x_j)\}$.
- ε-*neighborhoods* **graph** (gaussian kernel):
  $w_{ij} = max\{0, \frac{\varepsilon-||x_i-x_j||^2}{|\varepsilon-||x_i-x_j||^2|} \cdot e^{-\frac{||x_i-x_j||^2}{t}}\}$,
- the *topic* **graph**:
  $$w_{ij} = \delta(i,j) \cdot cos(x_i, x_j)$$
  where $\delta(i,j) = 1$ only if a corpus category C can be found such that $x_i \in C$ and $x_j \in C$ and 0 otherwise.

Open Issues

Applicability of DR metrics to complex tasks

- Which applications and scenarios?
- Which training conditions?
- Which parameters (dimensionality, locality principles, ...)

Objectives

- Explore all these issues ...
- on a large scale
- Evaluate different types of embeddings

Experimental Set-Up

Corpora and Tasks

- Reuters-21578 and 20NewsGroup
- Task: Document Clustering
- Models: VSM, LSA, LPP, LSA+LPP

Data sets

Collection	Docs	Tok	Topics
Reuters 21578	19,675	18,349	30
20NewsGroups	18,828	21,500	20

Clustering Algorithm

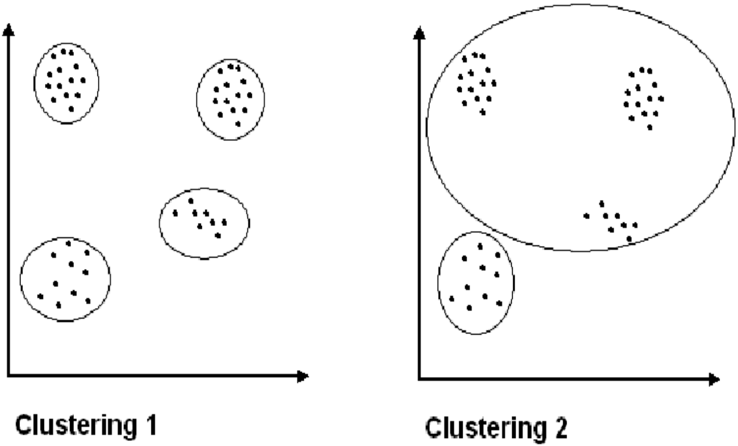
k-means

- Hard clustering algorithms fed with a fixed number of randomly chosen seeds (centroids)
- sensitive to the choice of k, and the seeding

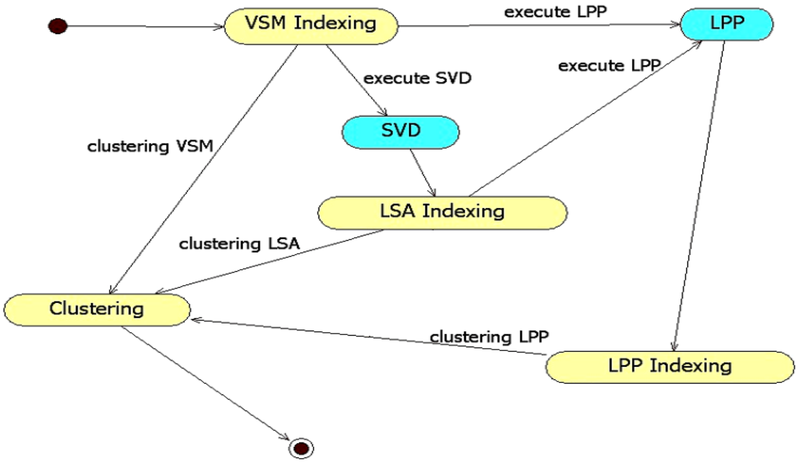
Adaptive variant ((Heyer et al., 1999))

- Agggregative clustering simiar to k-means with thresholds to increase flexibility
- Minimal infracluster similarity (activate *new seeds*)
- Maximal intra-cluster dissimilarity (activate *merge*)
- Maximal number of cluster members (activate *splits*)

Different settings



An overall view



Evaluation Metrics

NMI

Normalized mutual information, defined as follows:

$$NMI(T,C) = \frac{\sum_{t \in T, c \in C} p(t,c) log_2 \frac{p(t,c)}{p(t) \cdot p(c)}}{min(H(T), H(C))} \quad (1)$$

Accuracy

The accuracy AC is given by:

$$AC = \frac{\sum_{i=1}^{n} \delta(A_i, O_i)}{N} \quad (2)$$

where N is the total number of documents and $\delta(A_i, O_i)$ is 1 only if $A_i = O_i$ and 0 otherwise

Results

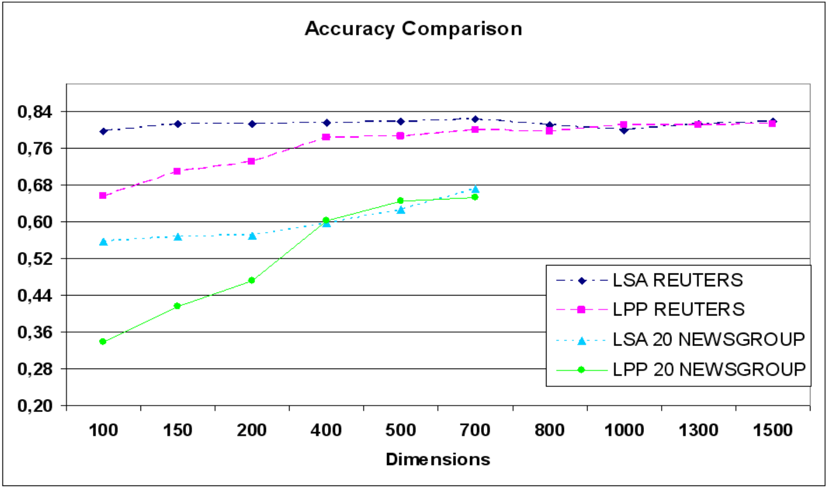
Topic Graph

METHOD	REUTERS	
	ACC	NMI
LSA	0.82	0.79
LPP	**0.94**	**0.99**

Table: Best LSA vs. upper bound LPP results based on the "*topic*" graph on Reuters.

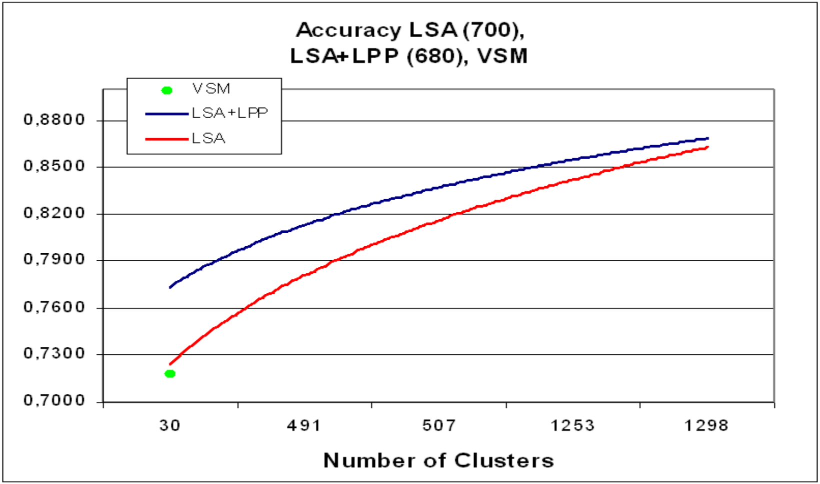
Results

Dimensionality reduction effect: LSA vs. LPP



Results

Clustering effect: LSA vs. LPP



Results

LSA vs. LSA+LPP (Reuters)

THR	LSA (700)		
	ACC	NMI	CLUSTERS
-1	0.72	0.61	30
0.2	0.82	0.79	507
0.4	**0.86**	**0.84**	1298

THR	LSA+LPP (LSA 700, LPP 680, ε=0.05)		
	ACC	NMI	CLUSTERS
-1	0.77	0.66	30
0.2	0.81	0.78	491
0.4	**0.86**	**0.84**	1253

Table: Performances on Reuters

Results

LSA vs. LSA+LPP (20Newsgroup)

THR	**LSA** (500)		
	ACC	NMI	CLUSTERS
-1	0.58	0.57	20
0.2	0.59	0.59	430
0.3	**0.63**	**0.64**	720

THR	**LSA+LPP** (LSA 500, LPP 480, ε=0.05)		
	ACC	NMI	CLUSTERS
-1	0.54	0.55	20
0.2	0.59	0.60	438
0.3	**0.62**	**0.64**	724

Table: Performances on 20Newsgroups

Conclusions

- This study shows that LSA and LPP improves the clustering accuracy even when much smaller number of features are employed

Conclusions

- This study shows that LSA and LPP improves the clustering accuracy even when much smaller number of features are employed
  - LPP alone is not competitive with LSA

Conclusions

- This study shows that LSA and LPP improves the clustering accuracy even when much smaller number of features are employed
  - LPP alone is not competitive with LSA
  - LPP can be succesfully combined with LSA

Conclusions

- This study shows that LSA and LPP improves the clustering accuracy even when much smaller number of features are employed
  - LPP alone is not competitive with LSA
  - LPP can be succesfully combined with LSA
- An interesting aspect explored here is the adoption of a priori knowledge in the design of the targeted locality principle
- The *topic graph* seems to provide the ideal space for clustering

Conclusions

Future Work

- Experiments LPP and LSA on other tasks, such as document classification and lexical disambiguation

Conclusions

Future Work

- Experiments LPP and LSA on other tasks, such as document classification and lexical disambiguation
- The definition of suitable adjacency graphs in LPP is an interesting research line, as several lexical learning tasks can be biased by existing lexical knowldge bases

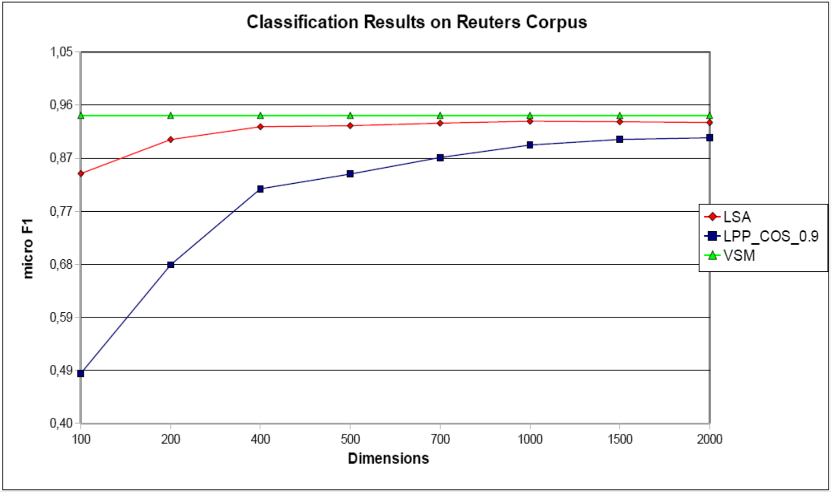
Conclusions

Future Work

- Experiments LPP and LSA on other tasks, such as document classification and lexical disambiguation
- The definition of suitable adjacency graphs in LPP is an interesting research line, as several lexical learning tasks can be biased by existing lexical knowldge bases
- Current work in modeling Framenet is on going

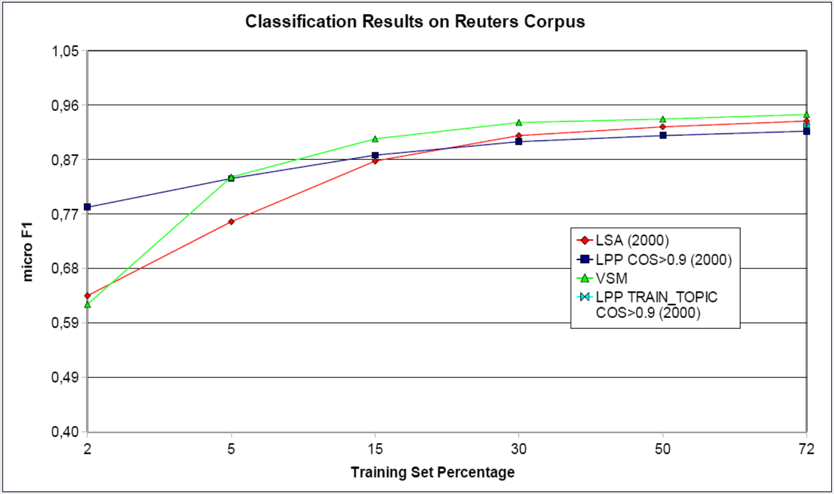
Results

Linear kernels for Text Classification (Reuters)



Results

Text Classification: Learning Rates



Thanks!