Text Representation strategies for Machine-Learning Categorization

D. A. Zighed and S. di Palma

Definition of Text Categorisation (TC)

- **TC = assignment of natural language texts t_i to one or more predefined categories C_s .**

- **Machine learning approach : automatically build a classifier F from a set Tr of n labelled texts t_i (the Training set).**
  => Routing : Implicit query defined by a set of positive and negatives examples.

- **Special case of Supervised Learning where the descriptive features are not given a priori.**

Applications of Text Categorisation

- **Automatic document indexing**
  - Boolean retrieval in digital libraries.
  - web pages categorization for hierarchy-based search systems.
- **Text routing**
  - find texts (or parts of text) that deserve a deeper examination in a complex system.
  - email filtering, personal information manager.
- **many other applications**
  - free questions in opinion surveys
  - automatic essay grading.
  - ….

Attribute-Value representation of the TC problem

	"Given"					To be predicted		
	T1	**…**	**Tj**		**Tk**	**C1**	**…**	**Cs**
text 1								
text 2			nij					cis
.								
text n								

T C

n_{ij} : value of the feature (attribute) Tj for the text i

c_{is} : value (0 or 1) of the label s for the text i

If each text belongs to only one category (non-overlapping case) , the problem can be tackled as a classical supervised learning problem with one class attribute **C** having s possible values.

Else : one classifier for each boolean class attribute C_s

Nature of the features

- **Using only statistical information about the typographic content of the texts**
  - words.
  - Statistical phrases (sometime also called Ngrams):
  - Ngrams : sequences of typographic symbols.

- **Using syntactic or semantic knowledge**
  - stems.
  - noun phrases.
  - key phrases.
  - semantic units.

Nature of the features
bag-of-words representations (1)

	T1	…	Tj		Tk	C
text 1						
.						
text i			nij			Ci
.						
text n						

T_j :

- boolean feature : absence/presence of the word j
- number of occurrences of the word j
- frequency

Nature of the features
bag-of-words representation (2)

Each feature represents a word

➨ **Advantages**

- intuitive and simple representation

➨ **drawbacks :**

- loss of all word order information
- high dimensionality of the problem : the set of potential features is made of all the words that appear at least one time in one of the texts.

Nature of the features
statistical phrase-based representation

Each feature represents a sequence of contiguous words

➨ **Expected advantage**

– phrases are more informative than single words

ex : “machine learning”, “world wide web”

➨ **Drawbacks :**

– greater number of potential features

– very low frequencies.

Nature of the features
Ngram-based representation (1)

Each feature represents a sequence of contiguous typographic symbols

ex : “great” => 3 trigrams : “gre”, “rea” and “eat”

➨ **Advantages**

- language-independent strategy
- applicable to any sequence : DNA, Ideograms, Music ...
- implicitly takes into account semantic or grammatical information

  ex : “drive” and “driven” have 3 trigrams in common

Nature of the features
Ngram-based representation (2)

- **Drawbacks of Ngram based representation:**
  - words meaning is lost => less intelligible models.
  - very “noisy” features

    ex : “drive” and “grive” have 3 trigrams in common.
  - choice of n.
  - High dimensionality of the problem (especially when n grows)

Nature of the features
Stems extraction (stemming) (1)

Each feature represents a stem.

ex : occurrences of "great" and "greater" are aggregated in the same feature "great"

➨ **Advantages**

- reduced number of features.
- reduced number of correlated features.
- simpler and more comprehensible classifiers.

Nature of the features
Stems extraction (stemming) (2)

- **Drawback : stemming is performed using knowledge-based systems. (eg M. Porter stemmer) =>**
  - language-dependent stemming systems.
  - computational cost
  - some stemming systems require a pre-processing tagging step.
    => still more time-consuming.

Nature of the features
Noun phrases and key phrases (1)

- **Noun Phrases : each feature represents a sequence of nouns and adjectives of length < n.**

- **Key Phrases : search for most informative sequences of words using a scoring function that takes into account :**
  - statistical information
    - number of words
    - position in the text
  - syntactic information
    - number of stems
    - presence or absence of stopwords ...

Nature of the features
Noun phrases and key phrases (2)

- **Expected advantages**
  - phrases are more informative than single words
  - simpler and more comprehensible classifiers.

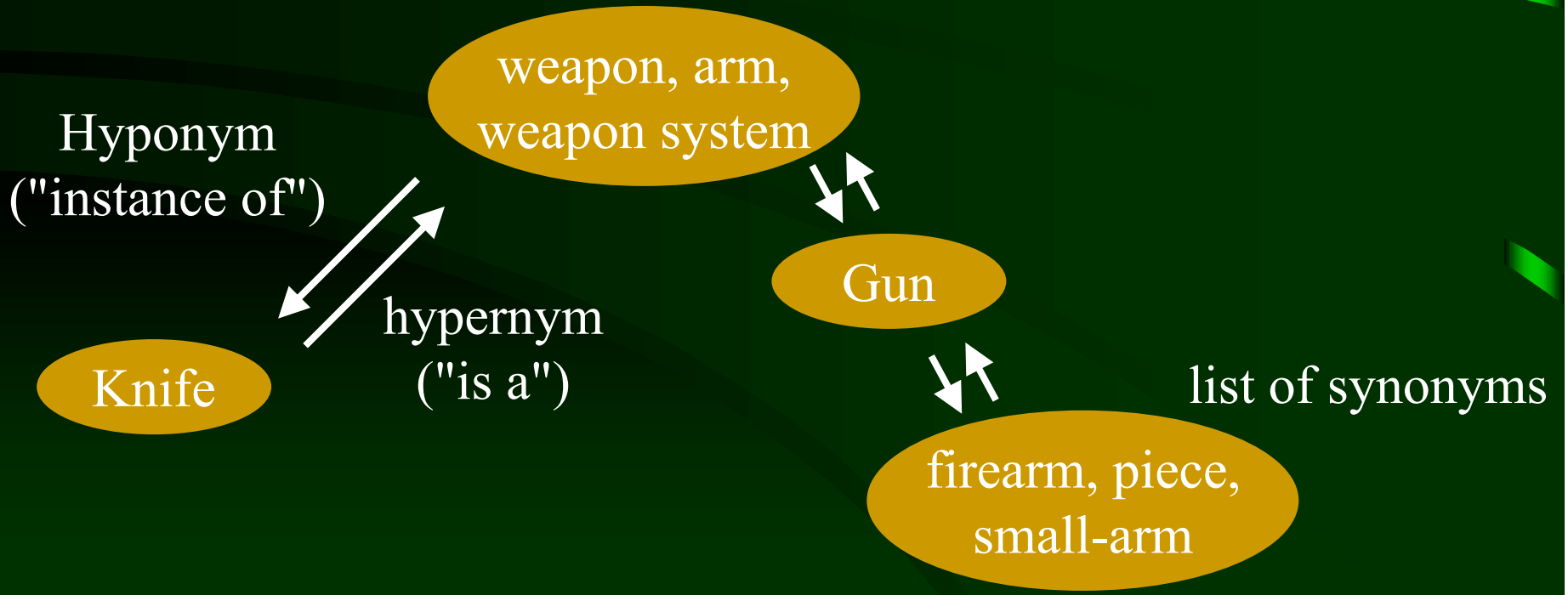
- **Drawbacks :**
  - language-dependent systems.
  - computational cost.
  - pre-processing tagging step required.
  - very low frequencies.

Nature of the features
semantic units (1)

Use of a semantic dictionary to extract features representing "word senses".

An idealised piece of the WordNet semantic hierarchy

(Sam Scott, 1998)



Nature of the features
semantic units (2)

- **Expected advantages**
  - simpler and more comprehensible classifiers.

- **Drawbacks :**
  - need for a semantic dictionary
  - language-dependent systems.
  - computational cost.
  - Polysemy => word sense disambiguation

Nature of the features simple but efficient ...

Recurrent discouraging result : sophisticated representation are most of the time useless as regards classification accuracy in TC, probably because of too low frequencies of phrases.

- S. Jones, 1981
- G. Salton, 1986, 1988
- M. Kee, 1981
- D. D. Lewis, 1992
- S. Scott et S. Matwin, 1998

➨**most systems use a representation based on stemmed words**

Specificity of TC as a Supervised Learning Problem (1)

- **High dimensionality**
  - several thousands of features
- **Sparsely distributed features**
  - most terms (word, words sequence, Ngram) occur only in a small fraction of the texts.
- **Imbalance in labels distributions.**
  - only a small number of texts belong to each category.
  - Moreover, the "non-relevant" category, in binary problems, refers to very heterogeneous texts

Specificity of TC as a Supervised Learning Problem (2)

- **Redundancy**
  - some terms are likely to occur together
    => correlated features
- **Synonymy**
  - occurrences of terms having the same meaning should be aggregated
- **Polysemy**
  - the same term can have several meanings
    => noisy features, necessity of word sense disambiguation

Dimension Reduction (DR) techniques

- **Feature selection and feature construction.**
  - Selection :

    the new set of features is a subset of the original one.
  - Construction

    the new features are derived from the original ones, most of the time through classification/clustering or linear combination.

- **Use of labels in DR**
  - no use of labels : unsupervised DR.
  - use of one label at a time : local supervised DR.

    one DR procedure (one texts representation) per label
  - use of all labels at the same time : global supervised DR.

Dimension Reduction (DR) techniques feature selection (FS)

mostly univariate filters

- univariate (≠ multivariate) : predictive features are assumed independent.
- filter (≠ wrapper): no feedback of learning results.

➨ **Unsupervised backward elimination of "uninteresting" terms.**

- stopwords removal
- frequency windowing

➨ **Supervised forward selection of "interesting" terms.**

- Aggressive statistical DR using a scoring function measuring the "importance"of a term for TC. This scoring function is often an association coefficient between the feature Tj representing a term and the class attribute C we want to predict.

Dimension Reduction techniques
FS using univariate scoring functions (1)

Most selection functions make only use of a binary form (presence/absence) of Tj => all the information is summarised by the following contingency table :

	Present	Absent	
C1	n1p		**n1**
…			
Cs		nsp	**ns**
	np	**na**	**n**

- n1p : number of texts belonging to the category C1 and containing the term Tj
- P(Cs) = ns / n = probability that a text belongs to Cs
- P(Cs/Tj) = nsa / na = probability that a text containing Tj belongs to Cs

binary problems

	Present	Absent	
Relevant	A	C	**nr**
Irrelevant	B	D	**ni**
	np	**na**	

Dimension Reduction (DR) techniques
FS using univariate scoring functions(2)

examples of scoring functions :

- mutual information
- Chi2
- odd-ratio
- cross-entropy / J-measure
- information gain
- ...

$$OddsRatio: \quad Odd(T_j) = \frac{\left(A/_{A+B}\right)+\left(B/_{A+B}\right)}{\left(C/_{C+D}\right)+\left(C/_{C+D}\right)}$$

$$Mutual\ Information: \quad MI(T_j) = \sum_i P(C_i).\log\left(\frac{P(C_i/T_j)}{P(T_j)}\right)$$

Dimension Reduction (DR) techniques
FS using univariate scoring functions(3)

➨ **Advantages**

- Low complexity.
- Practical efficiency.

➨ **Drawbacks :**

- Only use of presence/absence information.
- Terms independence assumption is unrealistic.
- Which scoring function should be used ?

Dimension Reduction techniques
feature Construction (FC) using Terms Clustering

Clustering of features => Aggregation of terms occurrences to create fewer new features

	T1	…	Tj	Tj'		Tk
text 1						
.						
text i			nij	nij'		
.						
text n						

	T1	…	Tnew		Tk
text 1					
.					
text i			nij+nij'		
.					
text n					

Dimension Reduction techniques
FC using Terms Clustering (2)

- **Semantic aggregation**
  - cf. use of semantic dictionaries

    (GUN or RIFLE) => WEAPON

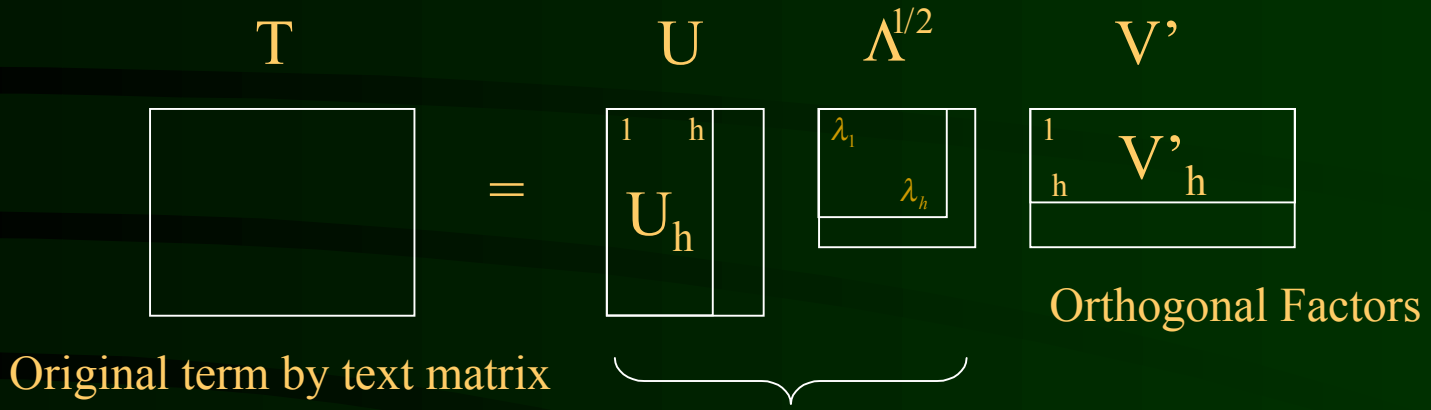
- **Statistical aggregation (example).**
  - Aggregate terms Tj and Tj' if the probability distributions P(C/Tj) and P(C/Tj') are close enough according to a particular distance measure (e.g. Kullback divergence).
  - Any other clustering method : K-Means, Hierarchical Clustering ...
  - Synonymy => use of overlapping clustering methods allowing a term to be classified in several clusters.

Dimension Reduction techniques
FC using Linear Algebra (1)

- **Applying linear algebra transformations to the original text-by-term matrix to capture most of its information in a lower-dimensional space.**
- **Taking into account terms correlations**
- **Traditional descriptive/exploratory techniques also used for feature construction : the new features are linear combinations of the original ones.**

Dimension Reduction techniques
FC using Linear Algebra (2)

Concept of Singular Value Decomposition (SVD)



$$\Psi_h = U_h \Lambda_h^{1/2}$$

the first factors (associated with the greatest singular values) represent linear combinations of the original terms that maximize the information retained from T in a lower dimension (according to the inertia criterion)

Dimension Reduction techniques
FC using Linear Algebra (3)

➨ **A semantic interpretation of SVD : Latent Semantic Indexing**

"factors may be thought of as artificial concepts; they represent extracted common meaning components of many different words and documents. […] Our aim […] is to represent terms […] in a way that escapes the unreliability, ambiguity and redundancy of individual terms as descriptors."

S. Deerwester et al.

➨ **Remark :**

the LSI decomposition can be applied "locally" to capture the structure of a particular subset of texts (relevant texts). Other texts are then projected as additional observations in the local LSI Space.

Dimension Reduction techniques
FC using Linear Algebra (4)

➨ **Advantages**

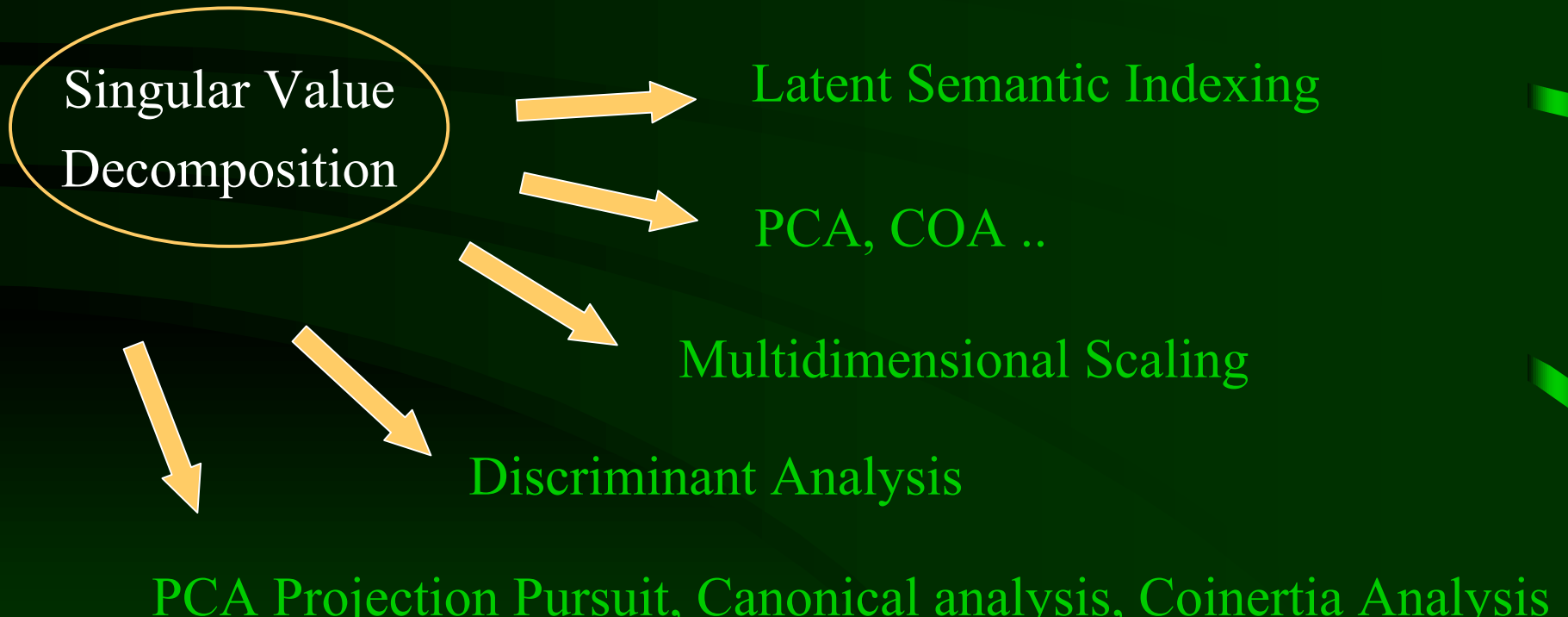
- Dimension reduction.
- Automatic detection of synonyms (decorrelation).
- Noise reduction by suppressing low inertia directions

➨ **Drawbacks :**

- Computational cost.
- The new description matrix is no more sparse
  => storage cost.

Dimension Reduction techniques
FC using Linear Algebra (5)

SVD is at the heart of many other dimension reduction techniques that have been (or could be …) used for feature extraction in text categorization tasks :



Dimension Reduction techniques
FC using Linear Algebra (6)

Principal Component Analysis (PCA) or KARHUNEN-LOEVE (KL) Transform

- Unsupervised multivariate DR technique.
- special case of SVD where original features have previously been centred.This centring procedure is not used in Information Retrieval applications to preserve the sparseness of the original data :
  - Lanczos SVD algorithm for sparse data : O(nk)
  - standard SVD algorithms : $O(n^2k)$.

Dimension Reduction techniques
FC using Linear Algebra (7)

Discriminant Analysis (DA)

- Multivariate DR technique supervised with respect to one predictive attribute.
- special case of SVD of the weighted barycentres of categories using W^{-1} (inverse of the intra-class variability matrix) as metrics.
- DA allows to extract only (number of categories -1) new features and is equivalent to multiple regression if there are only two categories.

Dimension Reduction techniques
FC using Linear Algebra (8)

Multivariate DR techniques supervised with respect to all the predictive attributes at the same time (overlapping case)

➨ **"multi-supervised Multivariate DR techniques" (MSMDR)...**

	T1	…	Tj		Tk	C1	…	Cs
text 1						1	1	0
.						0	1	0
text i			nij				1	1
.								
text n								

T C

➨ **Only one <u>supervised</u> DR to support the classification under all categories**

Dimension Reduction techniques
FC using Linear Algebra (9)

PCA Projection Pursuit (PP)

- Multivariate DR technique that tries to find "interesting" (usually as least Gaussian as possible) linear combinations of the original predictive attributes.
- Usually not used in IR applications for computational reasons.
- Some kind of PP can be achieved using standard linear algebra techniques : PCA with respect to a neighbourhood graph describing terms or texts proximities

  => allows MSMDR if the neighbourhood graph is derived from C.

  => also allows the use of semantic information in standard Multivariate DR techniques.

Dimension Reduction techniques
FC using Linear Algebra (10)

two-table coupling methods : MSMDR using direct analyses of the relationship between T and C

- Canonical Analysis : best known coupling method
  - **drawback** : inappropriate if the number of features of one table is greater than the number of texts (which is usually the case in TC).

- Most asymmetric method (allowing to predict C from T) have the same drawback.

- Co-Inertia Analysis

  symmetric (T and C play the same role) coupling method that does not have this drawback… to be tested soon in TC tasks ...

Conclusions

No universally appropriate text representation strategy

- **representation strategies must take into account :**
  - the nature of the TC task
  - the nature of the classifier that will be used in the learning step.

- **Need for text data-mining tools allowing to test easily many different learning algorithms and text representation strategies.**