Vectorial Representations: Parte 2

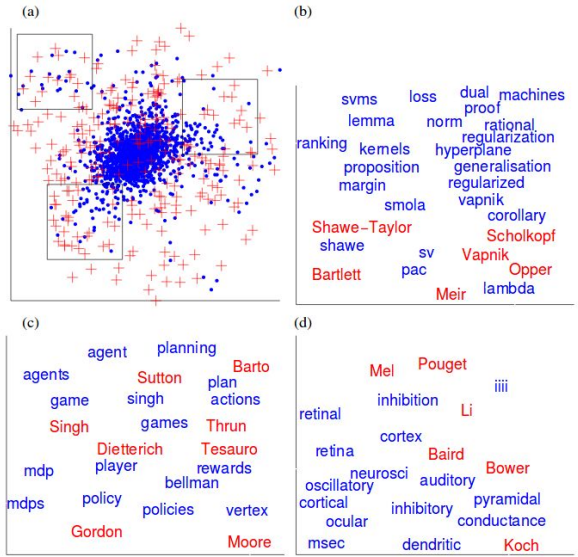
Word Vectors

Dr. Adrián Pastor López Monroy
Investigador

pastor.lopez@cimat.mx
https://www.cimat.mx/es/adrián-pastor-lópez-monroy

Centro de Investigación en Matemáticas, A.C. (CIMAT)

CIMAT



Globerson, A., Chechik, G., Pereira, F., & Tishby, N. (2007). Euclidean embedding of co-occurrence data. *Journal of Machine Learning Research*, *8*(Oct), 2265-2295.

Some slides are adapted from Diyi Yang, YuliaTsvetkovat (CMU) and Noah Smith at (UW): https://www.cc.gatech.edu/classes/AY2021/cs7650_fall/
https://drive.google.com/file/d/13GI64gjmaEMRBc3Qbh5DvkIIOisrD2eS/view
Other slides are adapted from: https://ccc.inaoep.mx/~mmontesg/cursos/Bogota/BeyondWords-UNAL2016.pdf

Outline of Word Vectors

- **Final Remarks BoW**
- **Distributional Terms Representations**
  - WordNet
  - DOR
  - TCOR
  - Random Indexing
  - Concise Semantic Analysis
  - Handcrafted Document Representations
- **Distributed Representations**
  - Latent Semantic Analysis (PCA for Text Mining)
  - Neural Word Embeddings (Word2Vec)

⋮

1) Ocurrencias
2) Co-ocurrencias

Automatica

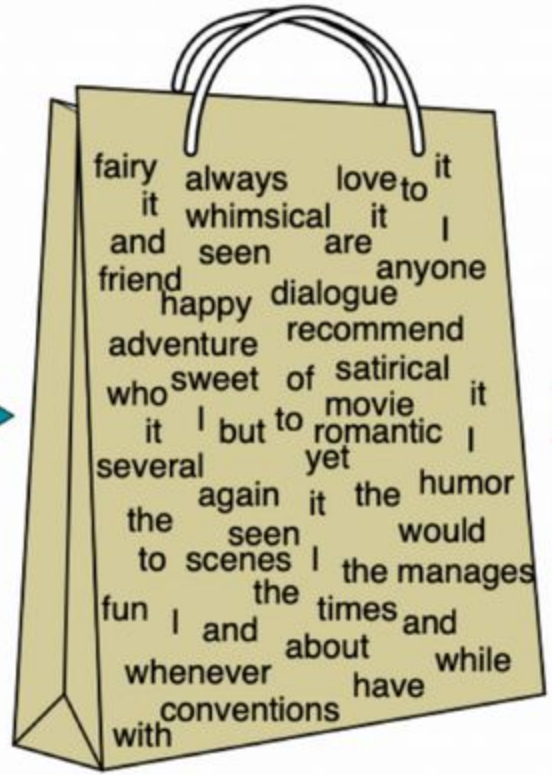
3) Esquemas de pesado

Dimensiones Interpretables

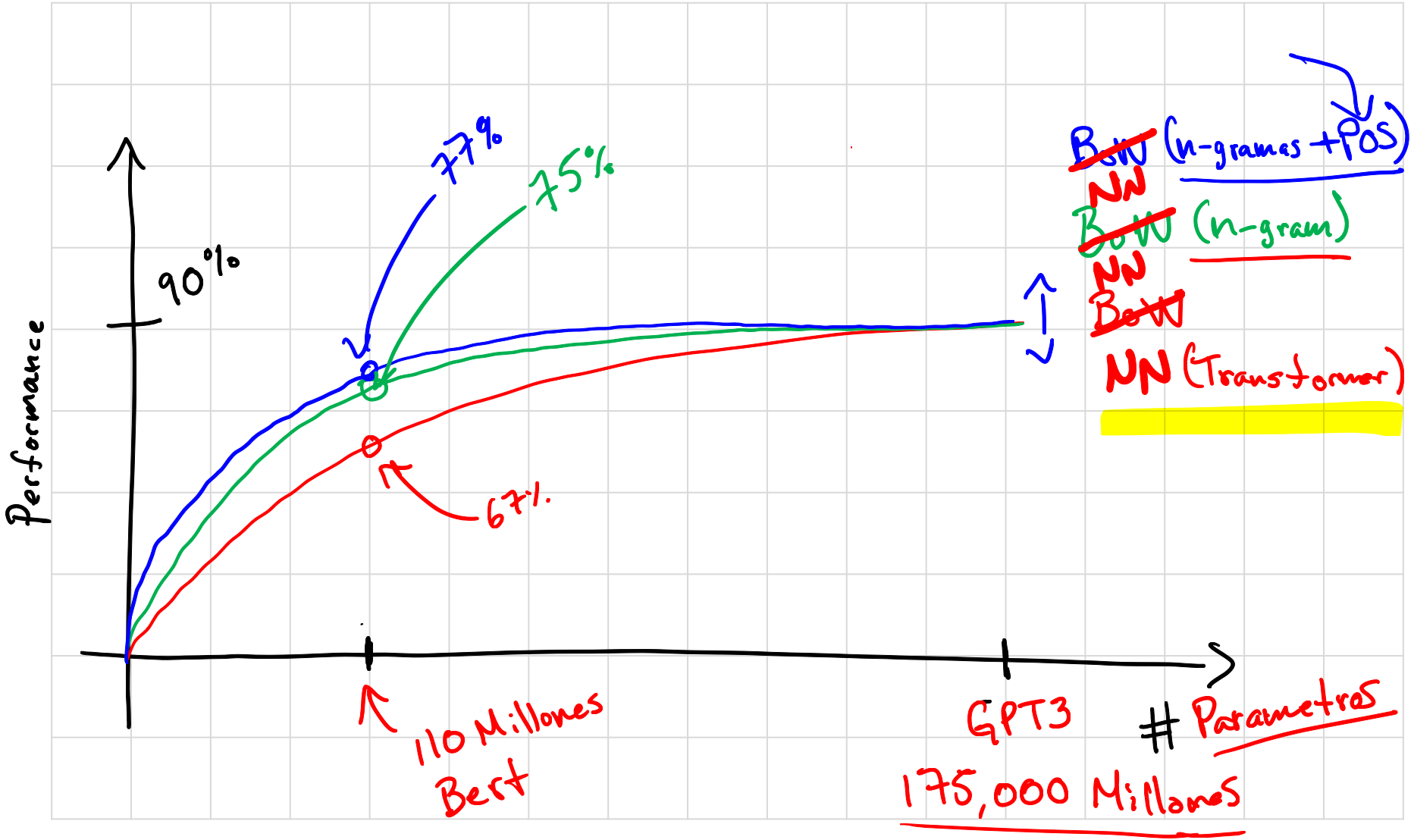
↳ Encajes
Incrostaciones

Bag-of-Words

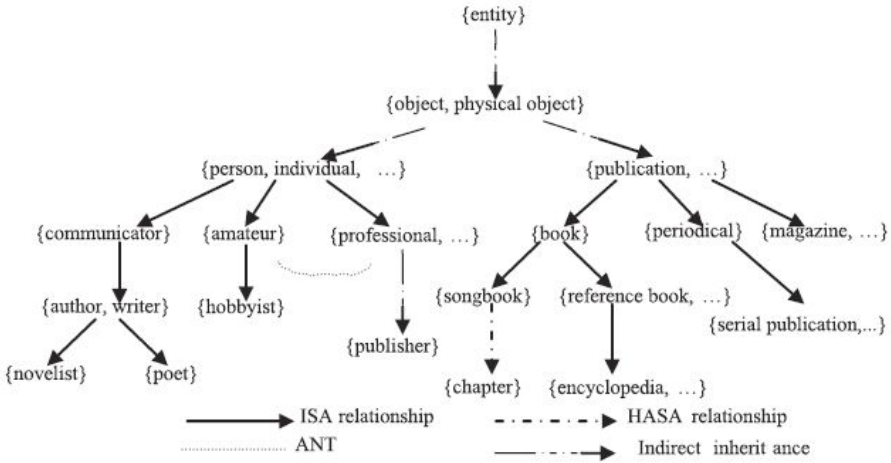
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

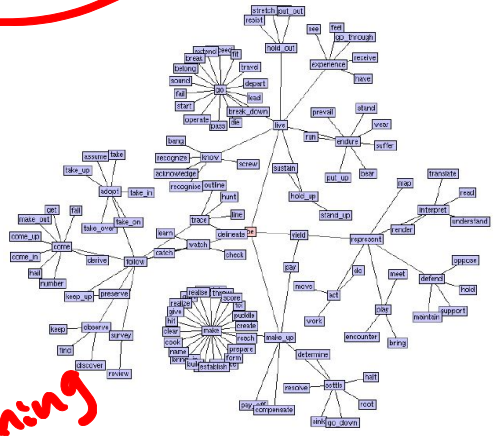


it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…


Performance
90%
77%
75%
67%
110 Millones
Bert
GPT3
175,000 Millones
Parametros
BoW (n-grames + POS)
NN
BoW (n-gram)
NN
BoW
NN (Transformer)

WordNet and More Corpora in NLTK





Deep Learning

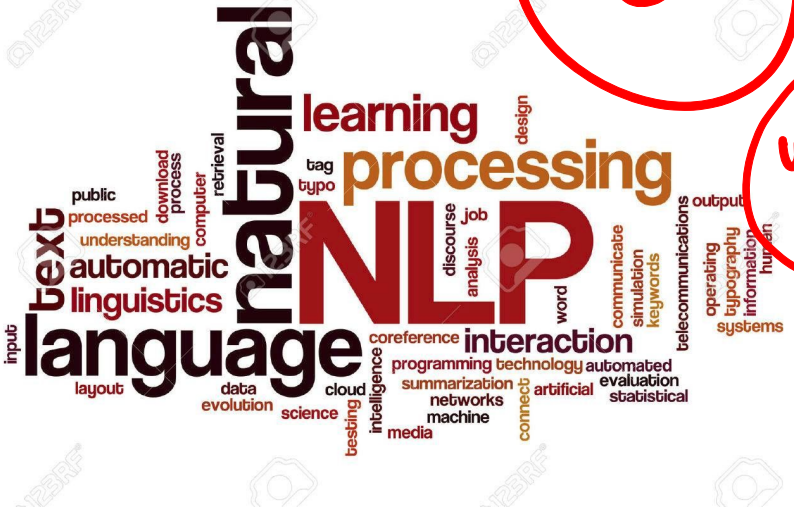
Reminder of Problems of BoW

- BoW ignores all semantic information; it simply looks at the surface word forms
  - Polysemy and synonymy are big problems
- BoW tend to produce very sparse representations, since terms commonly occur in just a small subset of the documents (difficult to find patterns)
  - This problem is amplified by lack of training texts and by the shortness of the documents to be classified.
- We need representations at concept level!
  - What if we do a finer analysis to built word vectors?
  - How to build Document Vectors from this? Advantages/Disadvantages
    - Text Classification is not the only application (e.g., Topic Analysis, Summarization, Translation, etc.)
    - **Bag-of-Concepts**

Ahora sí Word Vectors: Semántica Distribucional

"You shall know a word by the company it keeps"

John Rupert Firth, 1957



6

Word2Vec

Versión Español :-) "Dime con quién andas y ..."

Bag-of-Concepts

- Addresses the deficiencies of the BoW by considering the **relations between document terms.**
- BoC representations are based on the intuition that the meaning of a document can be considered as the **union of the meanings of their terms**.
- The meaning of terms is related to their usage; it is captured by their distributional representation **into a vector**.
  - Document occurrence representation (DOR)
  - Term co-occurrence representation (TCOR)
- Dimensions in word vectors usually are interpretable, but dimensions in document vectors may not.

Alberto Lavelli, Fabrizio Sebastiani, and Roberto Zanoli. Distributional term representations: an experimental comparison. *Thirteenth ACM international conference on Information and knowledge management* (CIKM '04). New York, NY, USA, 2004

Document Occurrence Representation (DOR)

- DOR representation is based on the idea that the semantics of a term may be view as a function of the bag of documents in which the term occurs.
  - Each document being an independent feature
- Terms are represented as vectors in the space of documents
  - Two terms are related if they show similar distributions across the documents

Terms Representation

	d_1	d_2	...	d_n
t_1				
t_2				
:		$w_{i,j}$		
t_m				

Short-Text Classification

The Key usually are in Term Weighting:

Short Text Classification Scenario

DOR

	d_1	d_2	...	d_n
t_1				
t_2				
:		$w_{i,j}$		
t_m				

$$w_{k,j} = df(d_k, t_j) \cdot \log \frac{|T|}{N_k}$$

$$df(d_k, t_j) = \begin{cases} 1 + \log(\#(d_k, t_j)) & if(\#(d_k, t_j) > 0) \\ 0 & otherwise \end{cases}$$

- DOR is a dual version of the BoW representation, therefore:
  - The more frequently t_i occurs in d_j, the more important is d_j for characterizing the semantics of t_i
  - The more distinct the words d_j contains, the smaller its contribution to characterizing the semantics of t_i

Representing Documents using DOR

$$\vec{d_i} = \sum_{t_j \in d_i} \alpha_{t_j} \vec{t_j}$$

DOR is a word representation, not a document representation.

Representation of documents is obtained by the weighted **sum of the vectors** from their terms.

MEX vs COL

$d_1 \ldots d_{1000}$ $d_{1001} \ldots d_{2000}$

$$d_i^{dtr} = \sum_{t_j \in d_i} \alpha_{t_j} \cdot w_{t_j}$$

Word representation

Word-Document matrix

	d_1	d_2	...	d_{1000}
t_1				
t_2				
:		$w_{i,j}$		
t_m				

$\ldots d_{1001} \ldots d_{2000}$

ahorita

chilango

⋮

Parse

SUM

Document representation

Document-Document matrix

	d_1	d_2	d_3	d_{1000}
d_1				
d_2				
d_{1000}		$w_{i,j}$		
d_{1001}				

$d_{1001} \ldots d_{2000}$

⋮

d_{2000}

Term CO-occurrence Representation (TCOR)

- In TCOR, the meaning of a term is conveyed by the terms commonly co-occurring with it; i.e. terms are represented by the **terms occurring in their context**
- Terms are represented as vectors in the space of terms (vocabulary of the collection)
- Two terms are related if they show similar co-occurring distributions with the rest of the terms

Representation of terms

	t_1	t_2	...	t_m
t_1				
t_2				
:		$w_{i,j}$		
t_m				

The Key usually are in Term Weighting:
Short Text Classification Scenario

	t_1	t_2	...	t_m
t_1				
t_2				
:		$w_{i,j}$		
t_m				

t_j

$$w_{k,j} = tff(t_k, t_j) \cdot log\frac{|T|}{T_k}$$

$$tff(t_k, t_j) = \begin{cases} 1 + \log(\#(t_k, t_j)) & if(\#(t_k, t_j) > 0) \\ 0 & otherwise \end{cases}$$

- TCOR is the kind of representation traditionally used in WSD, therefore:
  - The more times t_k and t_j co-occur in, the more important t_k is for characterizing the semantics of t_j
  - The more distinct words t_k co-occurs with, the smaller its contribution for characterizing the semantics of t_j.

Representing documents using TCOR

- TCOR, such as DOR, are word representations, not document representations.
- Representation of documents is obtained by the weighted sum of the vectors from their terms.

$$d_i^{dtr} = \sum_{t_j \in d_i} \alpha_{t_j} \cdot w_{t_j}$$

Word representation
Word-Word matrix

	t_1	t_2	...	t_m
t_1				
t_2				
:				
t_m				

SUM

Document representation
Document-Word matrix

	t_1	t_2	...	t_m
d_1				
d_2				
:				
d_n				

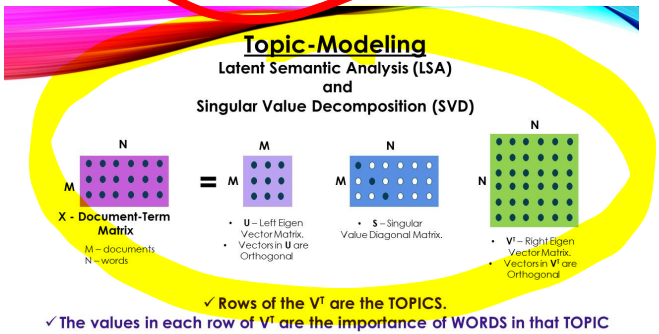
[BoW vs DOR vs TCOR] vs [LSA vs W2V]

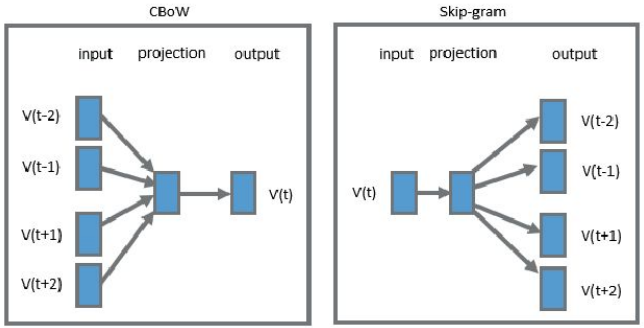
TCOR

	t_1	t_2	...	t_m
d_1				
d_2				
:				
d_n				

DOR

	d_1	d_2	...	d_n
d_1				
d_2				
:		w_i		
d_n				





https://www.datajango.com/topic-modeling-and-singular-value-decomposition-svd/

Distributional Semantics (DTRs)

- BOW
  - High dimensionality
  - Very sparse
- DOR
  - Lower dimensionality than BOW (d << w)
  - Not sparse
- TCOR
  - Same dimensionality than BOW (?, IG, PWMI)
  - Not sparse

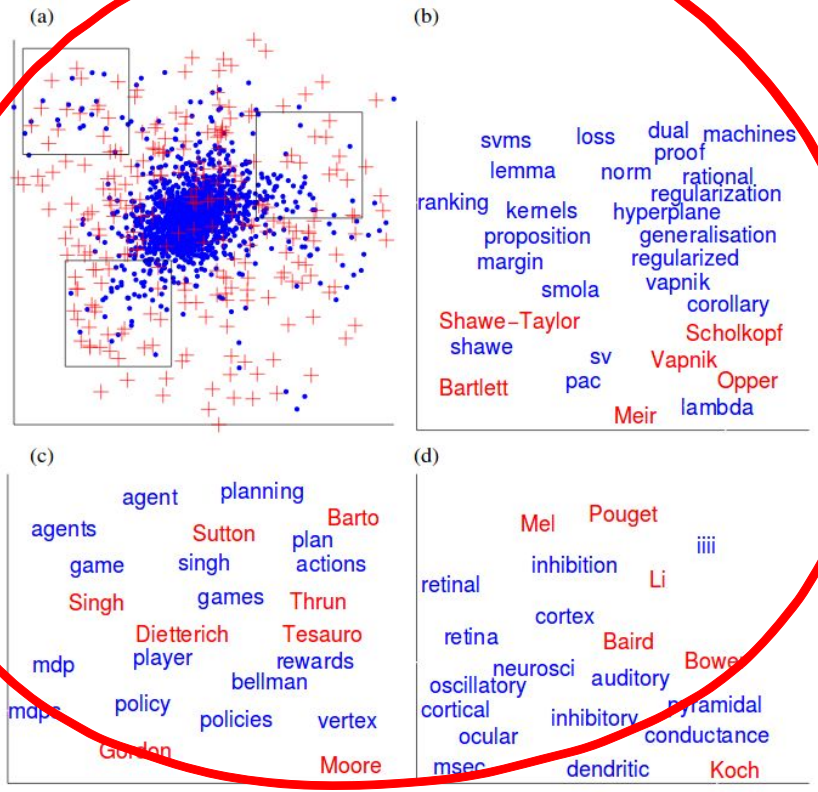
Distributed Semantics

- LSA
  - Low dimensional, Not sparse
  - Strong Mathematical concepts (SVD)
  - Very High Computational Cost: $O(n^3)$
- W2V
  - Low dimensional, Not sparse
  - Regression based
  - Variable window context
  - Very low computational cost

DOR and TCOR do a kind of expansion of the representations of documents

(DTR) Term Representation

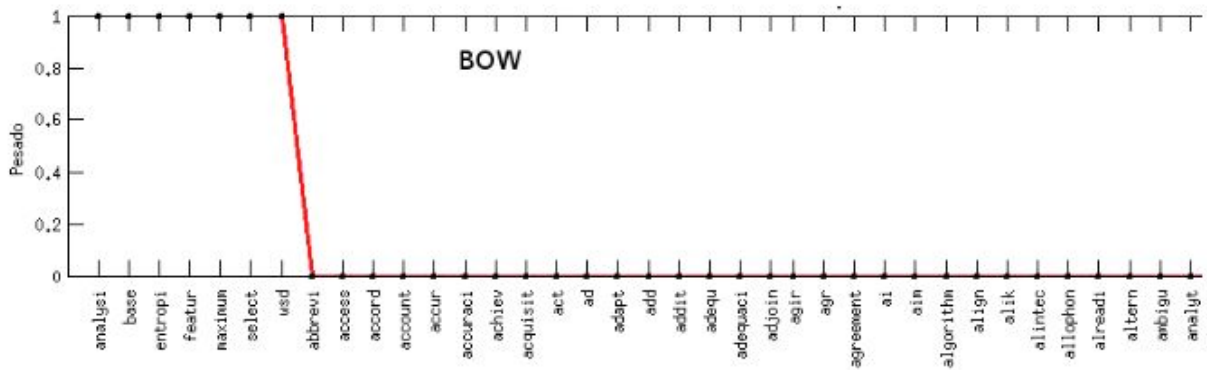
t-SNE



Globerson, A., Chechik, G., Pereira, F., & Tishby, N. (2007). Euclidean embedding of co-occurrence data. *Journal of Machine Learning Research*, *8*(Oct), 2265-2295.

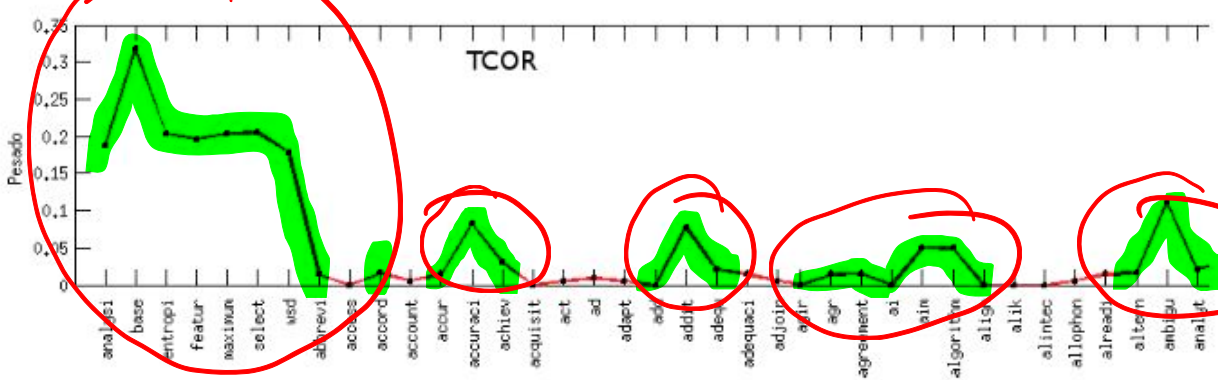
TCOR Representation of a paper Title

1



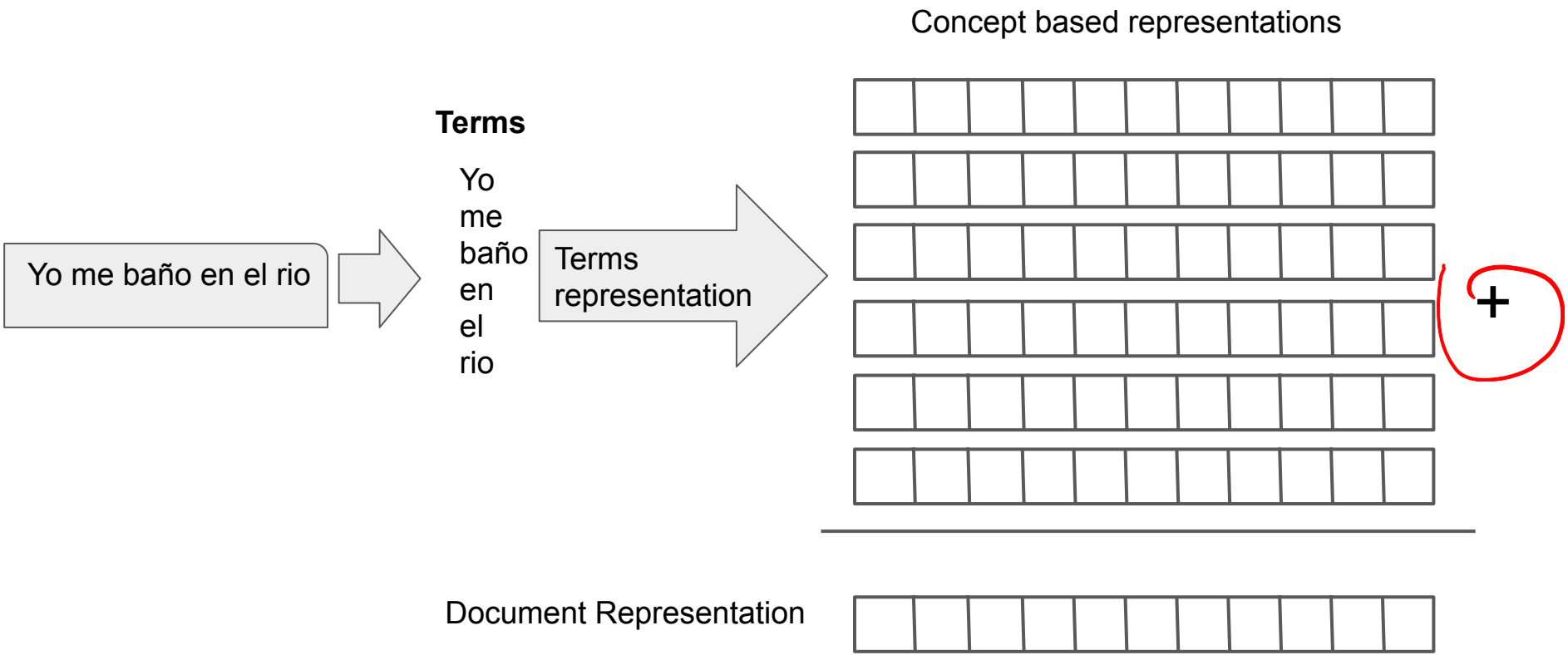
Feature selection analysis for maximun entropy based WSD

2



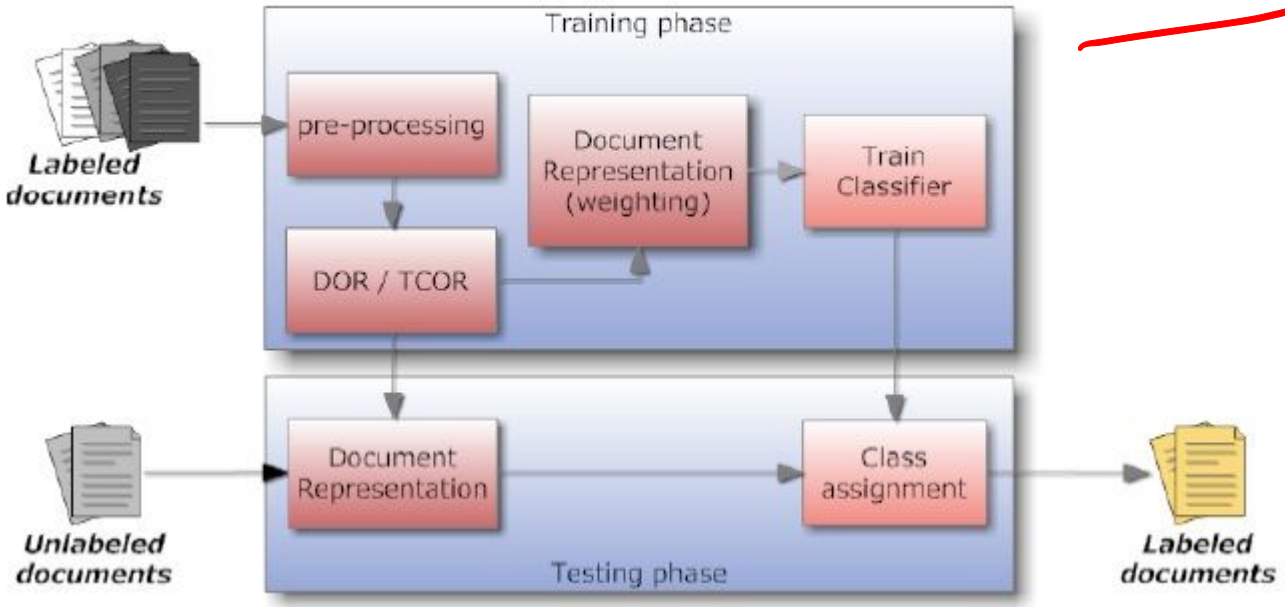
SVM

Simple Idea for Document building



DOR/TCOR for text classification

Random Indexing



Juan Manuel Cabrera, Hugo Jair Escalante, Manuel Montes-y-Gómez. Distributional term representations for short text categorization. *14th International Conference on Intelligent Text Processing and Computational Linguistics* (CICLING 2013). Samos, Greece, 2013.