DECSAI

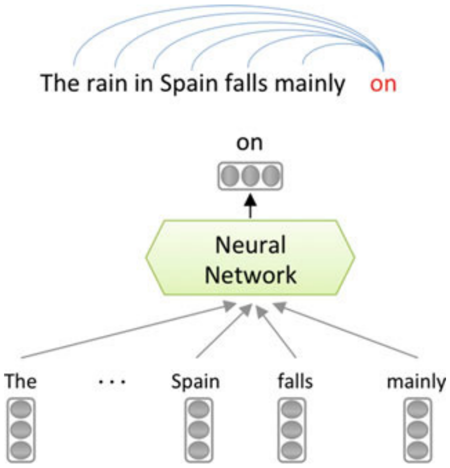
Departamento de Ciencias de la Computación e I.A.

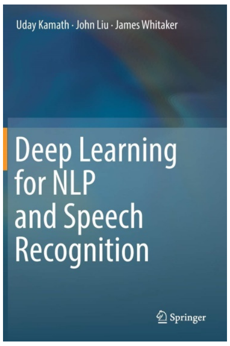
Universidad de Granada

Word embeddings

Fernando Berzal, berzal@acm.org

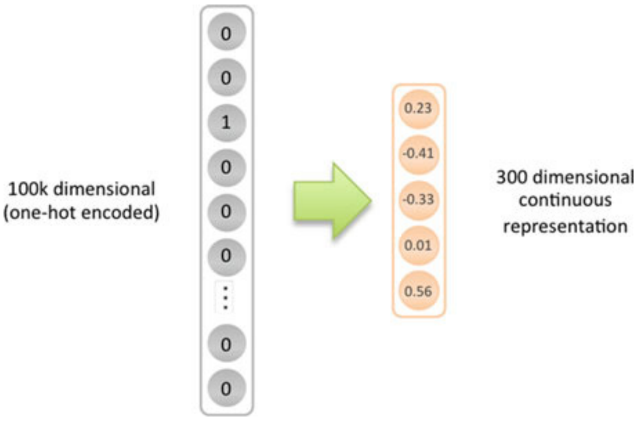
Modelo neuronal del lenguaje




Uday Kamath · John Liu · James Whitaker
Deep Learning for NLP and Speech Recognition
Springer

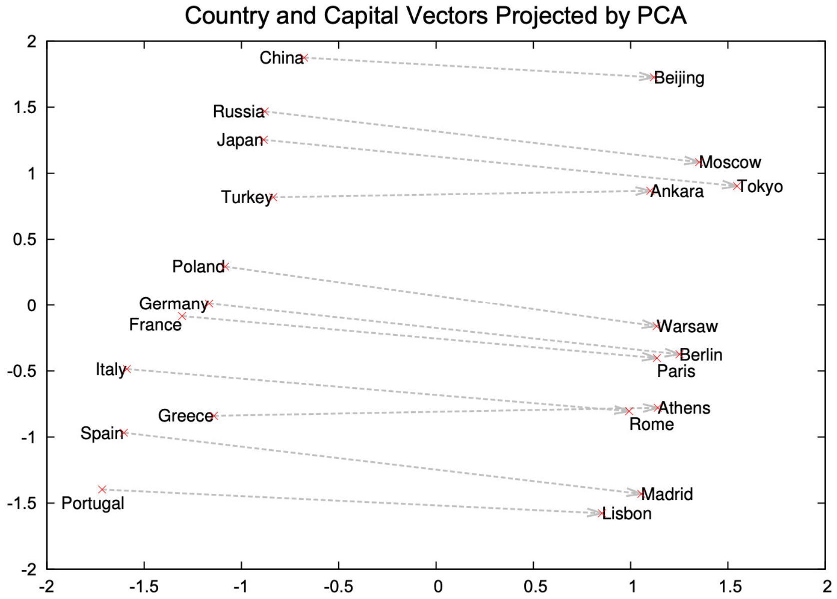
Joshua Bengio et al.
"A neural probabilistic language model". JMLR, 2003

Word embeddings



Ronan Collobert & Jason Weston:
“A Unified Architecture for Natural Language Processing:
Deep Neural Networks with Multitask Learning”. ICML’2008

Word embeddings



Mikolov et al.: “Distributed Representations of Words and Phrases and their Compositionality”, NIPS’2013

Word embeddings

- Relaciones semánticas

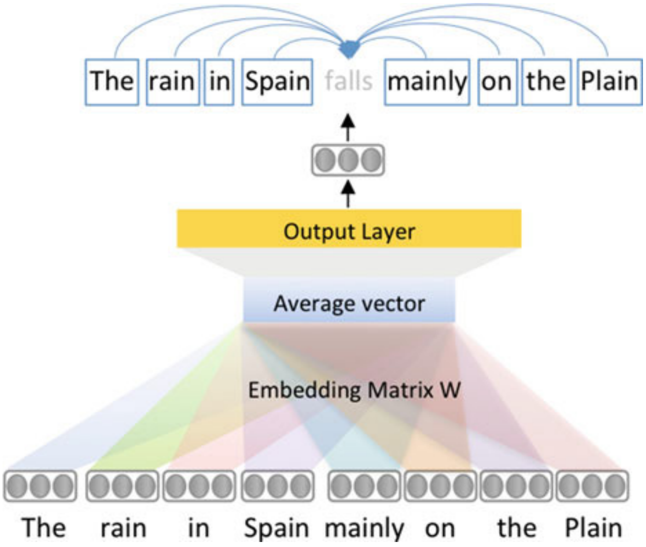
$$\mathbf{v}(\text{queen}) \approx \mathbf{v}(\text{king}) - \mathbf{v}(\text{man}) + \mathbf{v}(\text{woman})$$
$$\mathbf{v}(\text{Rome}) \approx \mathbf{v}(\text{Paris}) - \mathbf{v}(\text{France}) + \mathbf{v}(\text{Italy})$$
$$\mathbf{v}(\text{niece}) \approx \mathbf{v}(\text{nephew}) - \mathbf{v}(\text{brother}) + \mathbf{v}(\text{sister})$$
$$\mathbf{v}(\text{Cu}) \approx \mathbf{v}(\text{Zn}) - \mathbf{v}(\text{zinc}) + \mathbf{v}(\text{copper})$$

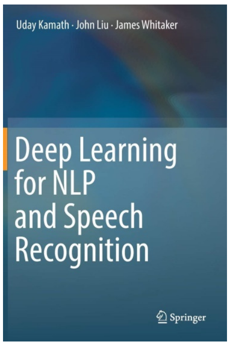
- Relaciones sintácticas

$$\mathbf{v}(\text{biggest}) \approx \mathbf{v}(\text{smallest}) - \mathbf{v}(\text{small}) + \mathbf{v}(\text{big})$$
$$\mathbf{v}(\text{thinking}) \approx \mathbf{v}(\text{read}) - \mathbf{v}(\text{reading}) + \mathbf{v}(\text{think})$$
$$\mathbf{v}(\text{mice}) \approx \mathbf{v}(\text{dollars}) - \mathbf{v}(\text{dollar}) + \mathbf{v}(\text{mouse})$$

word2vec

CBOW [Continuous bag of words]




Uday Kamath · John Liu · James Whitaker

Deep Learning for NLP and Speech Recognition

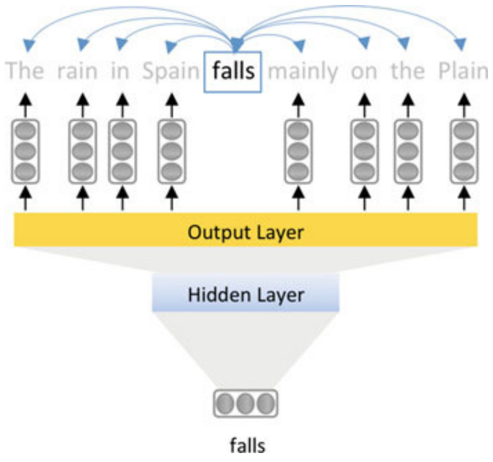
Springer

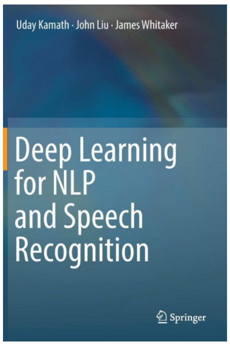
Context window = 4

Tomas Mikolov et al. "Distributed Representations of Words and Phrases and their Compositionality". NIPS'2013

word2vec

Skip-gram model





Uday Kamath · John Liu · James Whitaker

Deep Learning for NLP and Speech Recognition

Springer

Context window = 4

Tomas Mikolov et al. “Distributed Representations of Words and Phrases and their Compositionality”. NIPS’2013

GloVe

X_{ij} tabulate the number of times word j occurs in the context of word i.

$X_i = \sum_k X_{ik}$

$P_{ij} = P(j|i) = X_{ij}/X_i$

$w \in \mathbb{R}^d$ are word vectors

probe word

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}$$

co-relations between the word wi and wj

co-occurrence probabilities for the word wj and wk

$w_i{}^T \tilde{w}_k$ relate to (high probability if they are similar)

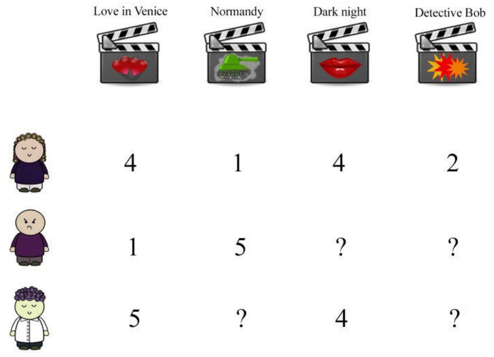
$$F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{P_{ik}}{P_{jk}}$$

$w_j{}^T \tilde{w}_k$

Jeffrey Pennington, Richard Socher & Christopher D. Manning: “GloVe: Global Vectors for Word Representation”. EMNLP’2014

GloVe

Se convierte en un problema de factorización de matrices (igual que en los sistemas de recomendación):



	Love in Venice	Normandy	Dark night	Detective Bob
	4	1	4	2
	1	5	?	?
	5	?	4	?

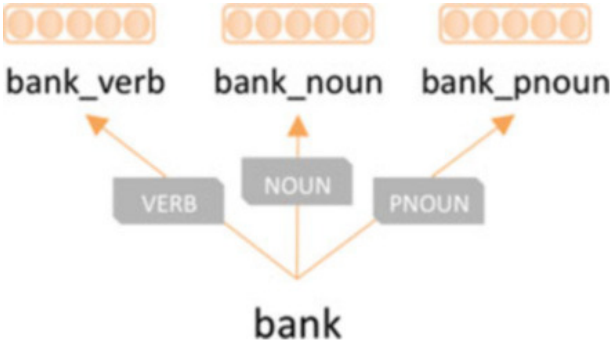
Jeffrey Pennington, Richard Socher & Christopher D. Manning: "GloVe: Global Vectors for Word Representation". EMNLP'2014

Limitaciones

- Palabras fuera del vocabulario [OOV]
- Antonimia
- Polisemia
- Sesgo (dependiendo del conjunto de entrenamiento)

$$\mathbf{v}(\text{nurse}) \approx \mathbf{v}(\text{doctor}) - \mathbf{v}(\text{father}) + \mathbf{v}(\text{mother})$$
$$\mathbf{v}(\text{Leroy}) \approx \mathbf{v}(\text{Brad}) - \mathbf{v}(\text{happy}) + \mathbf{v}(\text{angry})$$

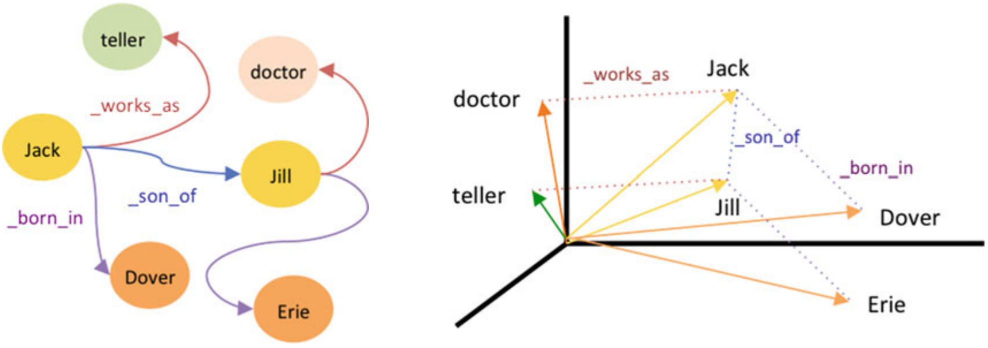
sense2vec



Andrew Trask, Phil Michalak & John Liu.
"sense2vec - A Fast and Accurate Method
for Word Sense Disambiguation in Neural Word Embeddings."
CoRR abs/1511.06388 (2015).

Más allá de las palabras…

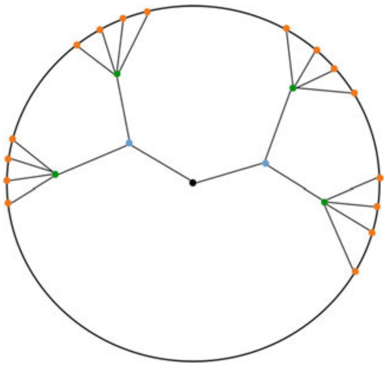
- Subword embeddings
- Sentence embeddings: Distributed memory [DM]
- Concept embeddings: RDF2Vec

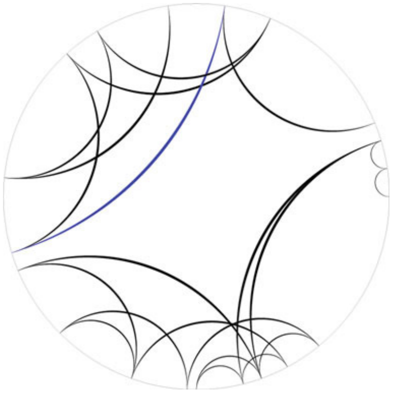


Más allá de las palabras…

- Gaussian embeddings: Word2Gauss
(distribuciones de probabilidad en lugar de vectores)

- Hyperbolic embeddings, a.k.a. Poincaré embeddings
(para relaciones jerárquicas)

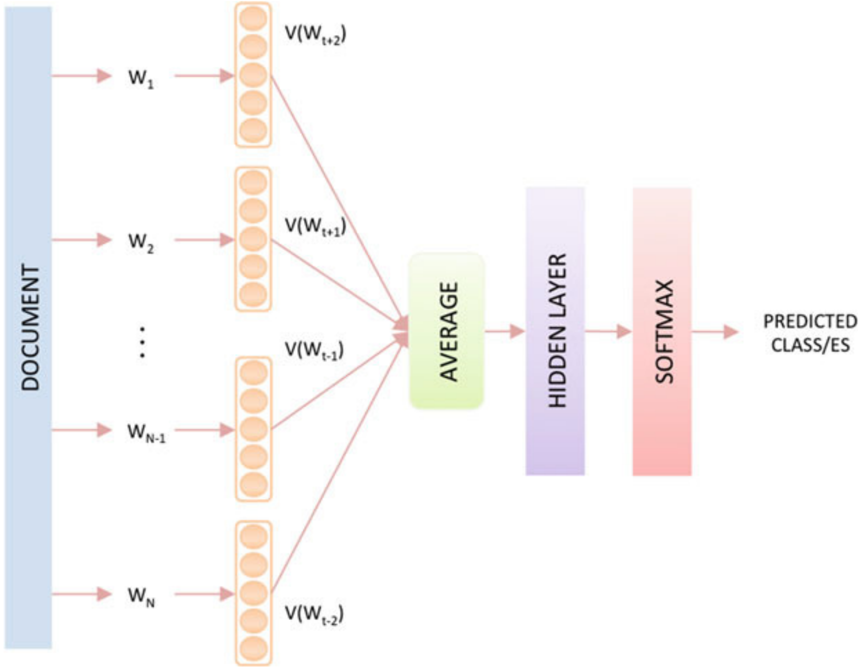




Aplicaciones

Clasificación de documentos

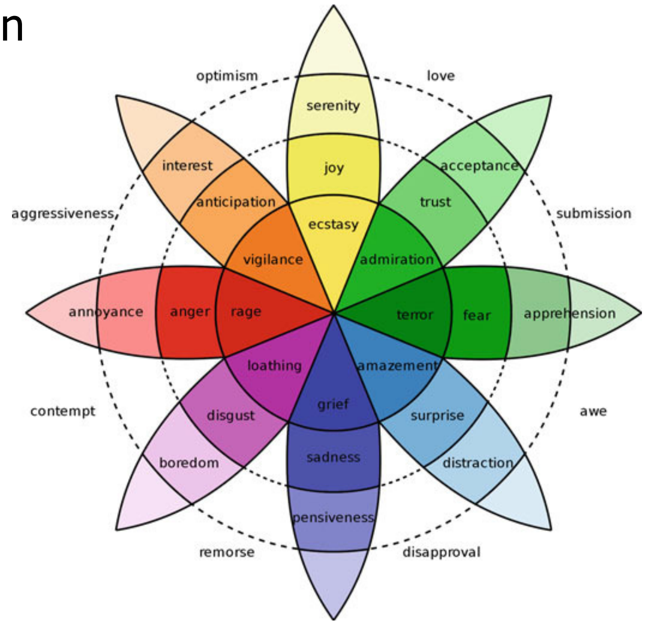
FastText



Aplicaciones

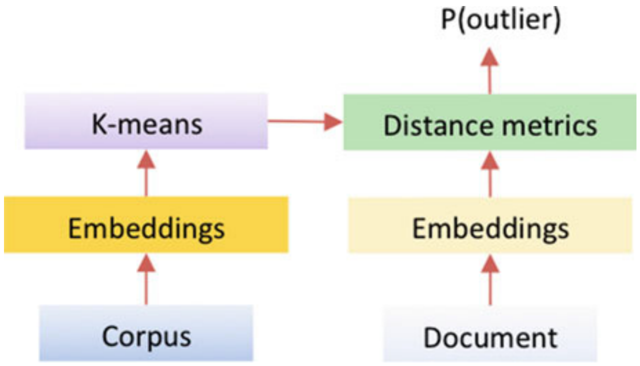
Clasificación de documentos

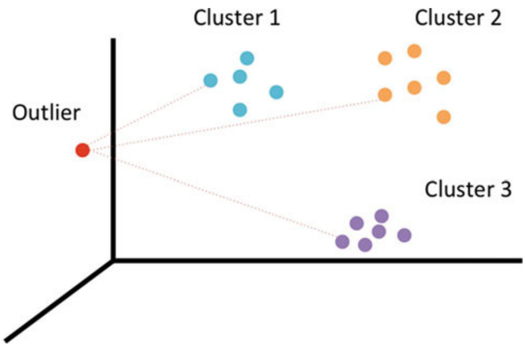
Sentiment classification



Aplicaciones

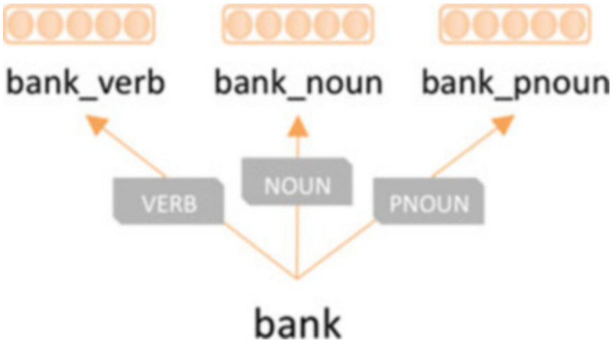
Detección de anomalías





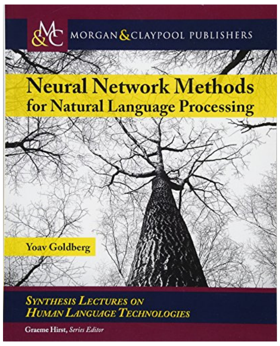
Aplicaciones

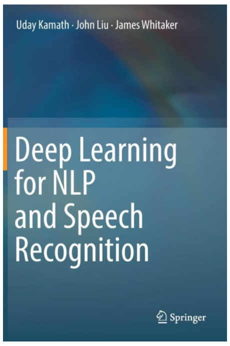
Word sense disambiguation [WSD]
p.ej. sense2vec POS tags



Bibliografía

Deep Learning & NLP


M&C MORGAN & CLAYPOOL PUBLISHERS
Neural Network Methods for Natural Language Processing
Yoav Goldberg
Synthesis Lectures on Human Language Technologies
Graeme Hirst, Series Editor


Uday Kamath · John Liu · James Whitaker
Deep Learning for NLP and Speech Recognition
Springer

- Yoav Goldberg:
  Neural Network Methods
  in Natural Language Processing
  Morgan & Claypool Publishers, 2017
  ISBN 1627052984
  https://doi.org/10.2200/S00762ED1V01Y201703HLT037

- Uday Kamath, John Liu & James Whitaker:
  Deep Learning for NLP and Speech Recognition
  Springer, 2019
  ISBN 3030145956
  http://link.springer.com/978-3-030-14595-8

Enlaces

Jonathan Hui: “NLP — Word Embedding & GloVe”, Medium, October 2019
https://jonathan-hui.medium.com/nlp-word-embedding-glove-5e7f523999f6