



Procesamiento de Lenguaje Natural

Embeddings

Mauricio Toledo-Acosta
mauricio.toledo@unison.mx

Departamento de Matemáticas
Universidad de Sonora



Procesamiento
de Lenguaje
Natural

Introducción e
Historia

Embeddings
de palabras

Embeddings
de
documentos

Section 1

Introducción e Historia



Sparsity problem

Procesamiento
de Lenguaje
Natural

Introducción e
Historia

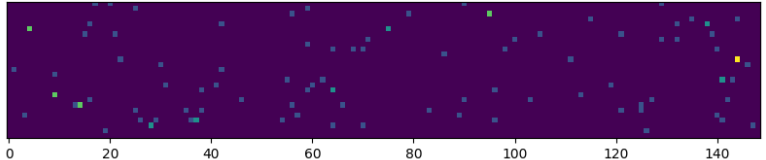
Embeddings
de palabras

Embeddings
de
documentos

Word2Vec



BOW





Bengio, 2003

Procesamiento
de Lenguaje
Natural

Introducción e
Historia

Embeddings
de palabras

Embeddings
de
documentos

Journal of Machine Learning Research 3 (2003) 1137–1155

Submitted 4/02; Published 2/03

A Neural Probabilistic Language Model

Yoshua Bengio
Réjean Ducharme
Pascal Vincent
Christian Jauvin

BENGIOY@IRO.UMONTREAL.CA
DUCHARME@IRO.UMONTREAL.CA
VINCENTP@IRO.UMONTREAL.CA
JAUVINC@IRO.UMONTREAL.CA

*Département d'Informatique et Recherche Opérationnelle
Centre de Recherche Mathématiques
Université de Montréal, Montréal, Québec, Canada*

Editors: Jaz Kandola, Thomas Hofmann, Tomaso Poggio and John Shawe-Taylor

The original paper



Antecedentes, 1991

Procesamiento
de Lenguaje
Natural

Introducción e
Historia

Embeddings
de palabras

Embeddings
de
documentos

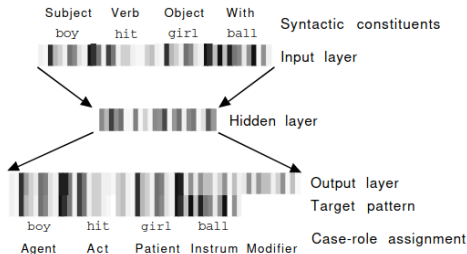


Figure 2: **Snapshot of basic FGREP simulation.** The input and output layers of the network are divided into assemblies, each holding one word representation at a time. Each unit in an input assembly is set to the activity value of the corresponding component in the lexicon entry. The input layer is fully connected to the hidden layer and the hidden layer to the output layer. Connection weights are omitted from the figure. If the network has successfully learned the task, each output assembly forms an activity pattern identical to the lexicon representation of the word filling that role. The correct role assignment is shown at the bottom of the display. This pattern forms the output target for the network. Grey-scale values from white to black are used in the figure to code the unit activities, which vary within the range [0,1].



Antecedentes, 1991

Procesamiento de Lenguaje Natural

Introducción e Historia

Embeddings de palabras

Embeddings de documentos

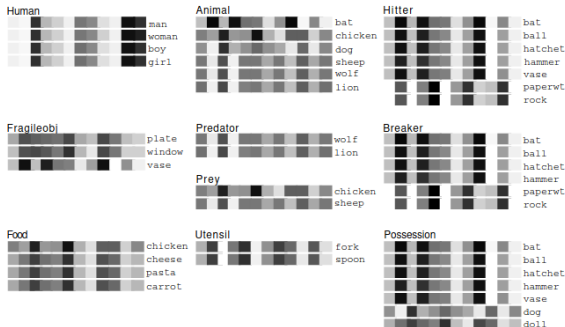


Figure 3: **Final representations.** The representations for the synonymous words {man, woman, boy, girl}, {fork, spoon}, {wolf, lion}, {plate, window}, {ball, hatchet, hammer}, {paperwt, rock} and {cheese, pasta, carrot} have become almost identical.



La hipótesis distributiva de la semántica

Procesamiento
de Lenguaje
Natural

Introducción e
Historia

Embeddings
de palabras

Embeddings
de
documentos

the.language.nerds

Me speaking English:

I have literally no idea what this word translates to in my native language but I've seen it being used in similar context so I'm just gonna use it here and pray that it does mean what I think it means.



¿Qué es el kimchi?

Procesamiento
de Lenguaje
Natural

Introducción e
Historia

Embeddings
de palabras

Embeddings
de
documentos

Contextos directos:

- Las recetas de **kimchi** cuentan con una gran cantidad de antioxidantes
- Para preparar **kimchi**, se corta la col china en trozos y se mezcla con sal
- Se puede servir el **kimchi** como acompañamiento del arroz, junto con la carne
- El **kimchi** se conserva en recipientes herméticos en el refrigerador



- Las **espinacas** se cortan en trozos y se saltean con ajo
- Sirve las **espinacas** como acompañamiento del arroz integral
- Las **acelgas** se conservan mejor en el cajón del refrigerador
- Para preparar las **acelgas**, se cortan las hojas y se separan los tallos
- Las **acelgas** son ricas en antioxidantes y fibra
- La **col** fermentada (chucrut) se prepara con sal y especias



Section 2

Embeddings de palabras



**He poured himself a cup of coffee
himself**

- | | |
|--------|---------------------------|
| input | <i>himself</i> |
| output | <i>He, poured, a, cup</i> |

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ↺ 🔍 ↻ 11/22



Ejemplo de entrenamiento Word2Vec

Procesamiento
de Lenguaje
Natural

Introducción e
Historia

Embeddings
de palabras

Embeddings
de
documentos

The quick brown fox jumps over the lazy dog and
runs away

The quick brown fox jumps over the



The quick brown fox jumps over the lazy dog and
runs away

quick brown fox jumps over the lazy



The quick brown fox jumps over the lazy dog and
runs away

brown fox jumps over the lazy dog



The quick brown fox jumps over the lazy dog and
runs away

fox jumps over the lazy dog and



The quick brown fox jumps over the lazy dog and
runs away

jumps over the lazy dog and runs



The quick brown fox jumps over the lazy dog and
runs away

over the lazy dog and runs away

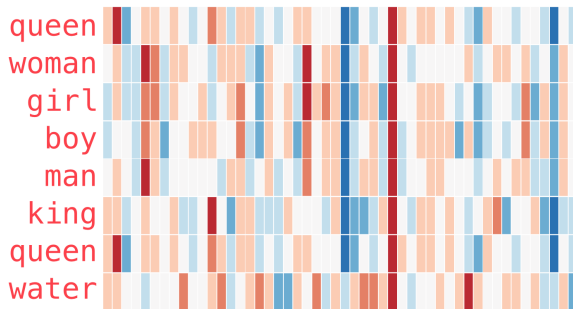


Procesamiento de Lenguaje Natural

Introducción e
Historia

Embeddings
de palabras

Embeddings
de
documentos





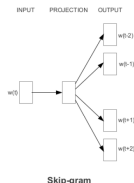
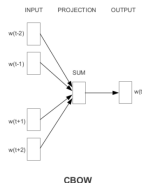
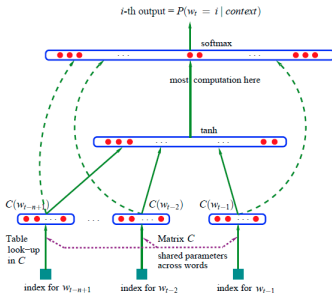
Diferencias entre los enfoques de NPLM y Word2Vec

Procesamiento
de Lenguaje
Natural

Introducción e
Historia

Embeddings
de palabras

Embeddings
de
documentos





- In contrast to word2vec, GloVe seeks to make explicit what word2vec does implicitly: Encoding meaning as vector offsets in an embedding space – seemingly only a serendipitous by-product of word2vec – is the specified goal of GloVe.
- There are no vectors for OOV words.

GloVe, Original paper



16/22



- FastText computes valid representations for OOV words (out-of-vocabulary) by taking the sum of its n -grams vectors.

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ↺ 🔍 ↻ 17/22



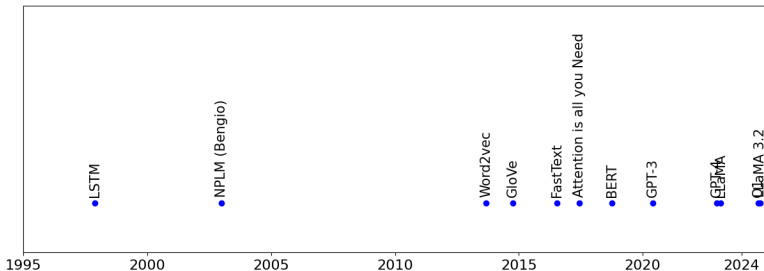
Timeline

Procesamiento de Lenguaje Natural

Introducción e Historia

Embeddings de palabras

Embeddings de documentos





Section 3

Embeddings de documentos



¿Cómo representamos documentos?

- Promedio de vectores de palabras (centroide).
- Promedio pesado de vectores de palabras.
- Usando redes neuronales.
- Usando embeddings de documentos:
 - doc2vec: Le and Mikolov in 2014 introduced the Doc2Vec algorithm, which usually outperforms such simple-averaging of Word2Vec vectors. The basic idea is: act as if a document has another vector, which contributes to all training predictions, and is updated like other word-vectors, but we will call it a doc-vector.
[Original paper Gensim's doc2vec](#)
 - Bert-based embeddings.



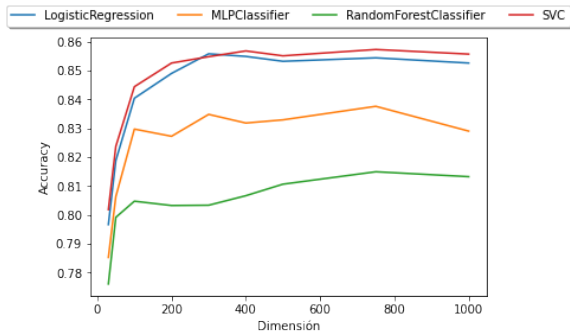
Efecto de la dimensión

Procesamiento
de Lenguaje
Natural

Introducción e
Historia

Embeddings
de palabras

Embeddings
de
documentos



La representación de cada documento está dada por el promedio de cada vector de word2vec.



¿Aún son vigentes estos modelos?

Procesamiento de Lenguaje Natural

Introducción e Historia

Embeddings de palabras

Embeddings de documentos

