Aplicación de Procesamiento de Lenguaje Natural en la Ciencia

Gustavo Vazquez

gustavo.vazquez@ucu.edu.uy tw: @VazquezGust

Universidad Católica del Uruguay

Depto. de Ciencias de la Computación - Facultad de Ingeniería y Tecnologías

- La ciencia genera un gran volumen de información (típicamente no estructurada)

- Algunos ejemplos:
  - Diseño de Materiales/Diseño de Drogas (Cheminformatics)
  - Biología (Bioinformática)


COMMUNICATIONS

ORGANIC LETTERS 2003 Vol. 5, No. 14 2453–2455

Mild Method for Ullmann Coupling Reaction of Amines and Aryl Halides

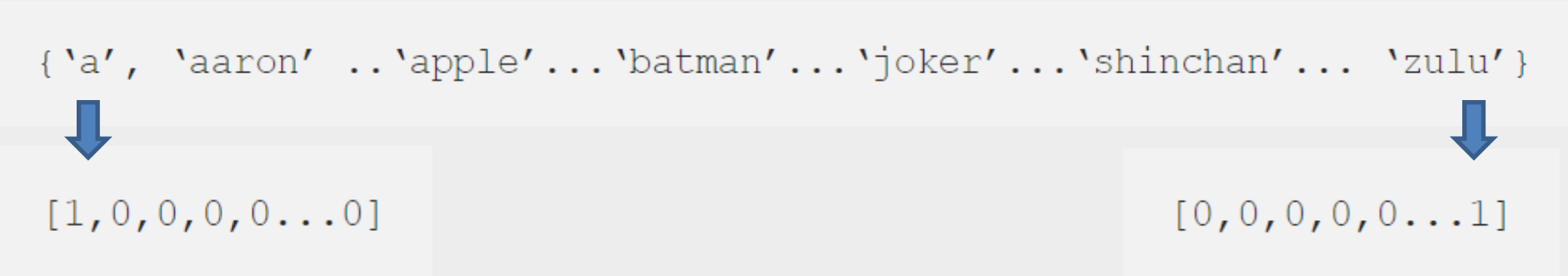
Dawei Ma,*† Qian Cai,† and Hui Zhang†

ABSTRACT

the Reaction of 1H-Pyrazol-5-amines with 4,5-thiazolium Chloride: A Route to thiazoles and Pyrazolo[3,4-d]thiazoles

Manos J. Manolis, and Panayiotis A. Koutentis*

- One-Hot encoding



- Problemas de esta representación
  - Sparse
  - No preserva aspectos semánticos

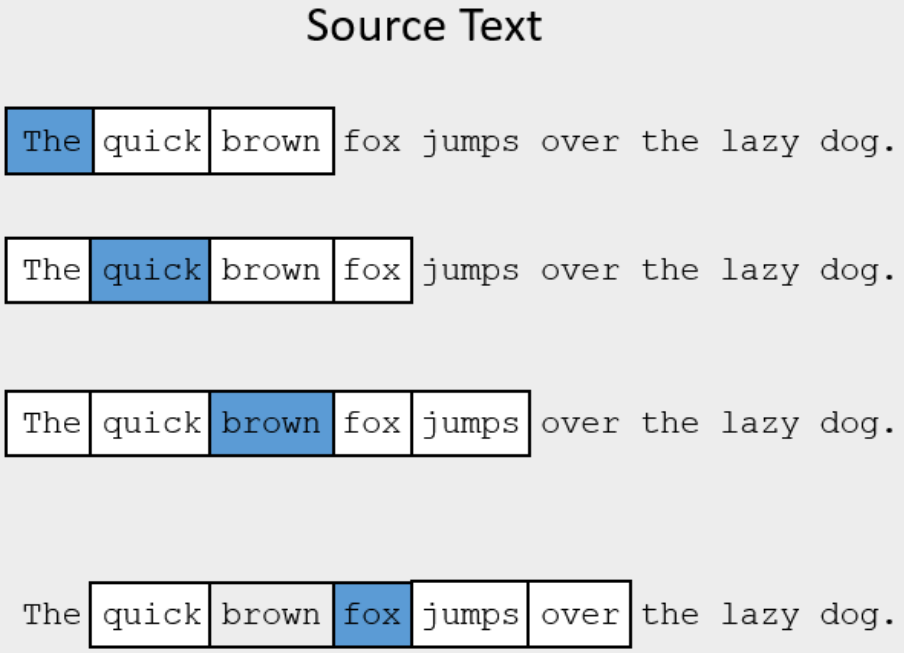
- Word Vectors: qué sucedería si…

```
word2vec('Batman') = [0.9, 0.8, 0.2]

word2vec('Joker') = [0.8, 0.3, 0.1]

word2vec('Spiderman') = [0.2, .9, 0.8]

word2vec('Thanos') = [0.3, 0.1, 0.9]
```

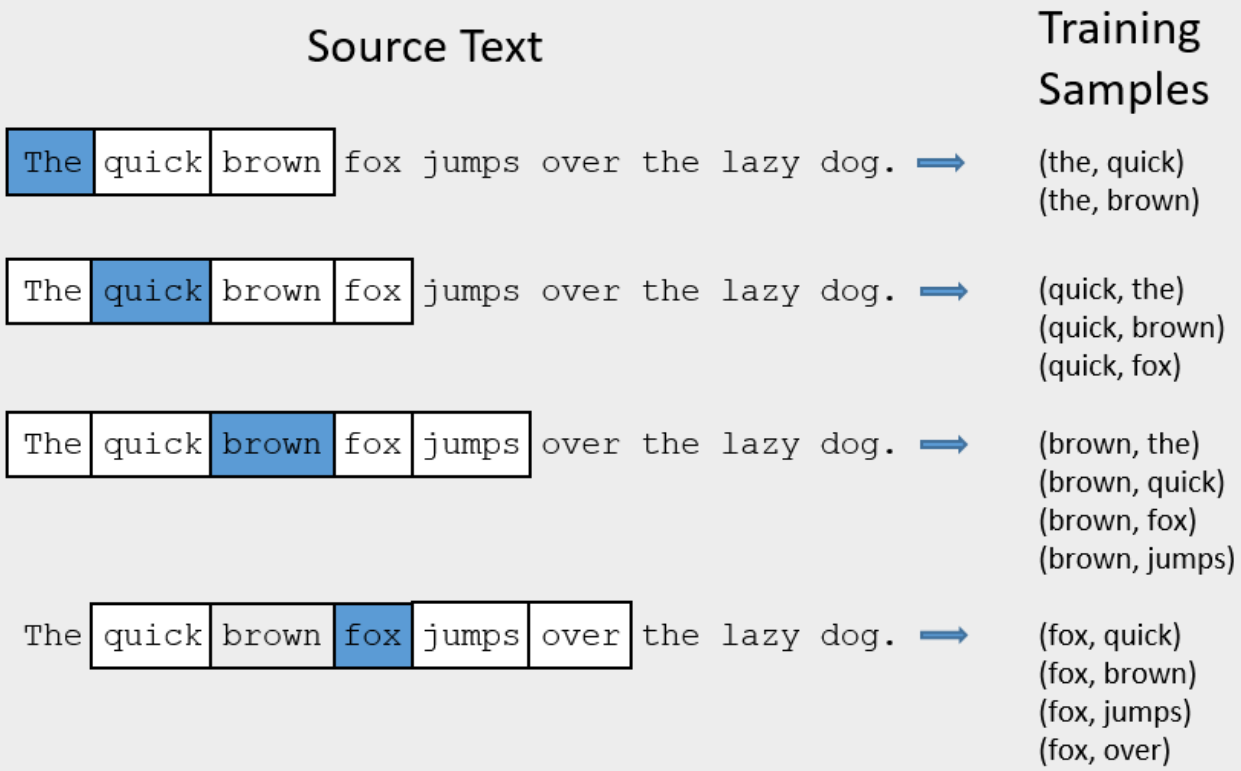
- Ventajas
  - Los vectores pueden codificar características / información semántica
  - Reducción de la dimensión

¿Cómo obtenemos esta representación?

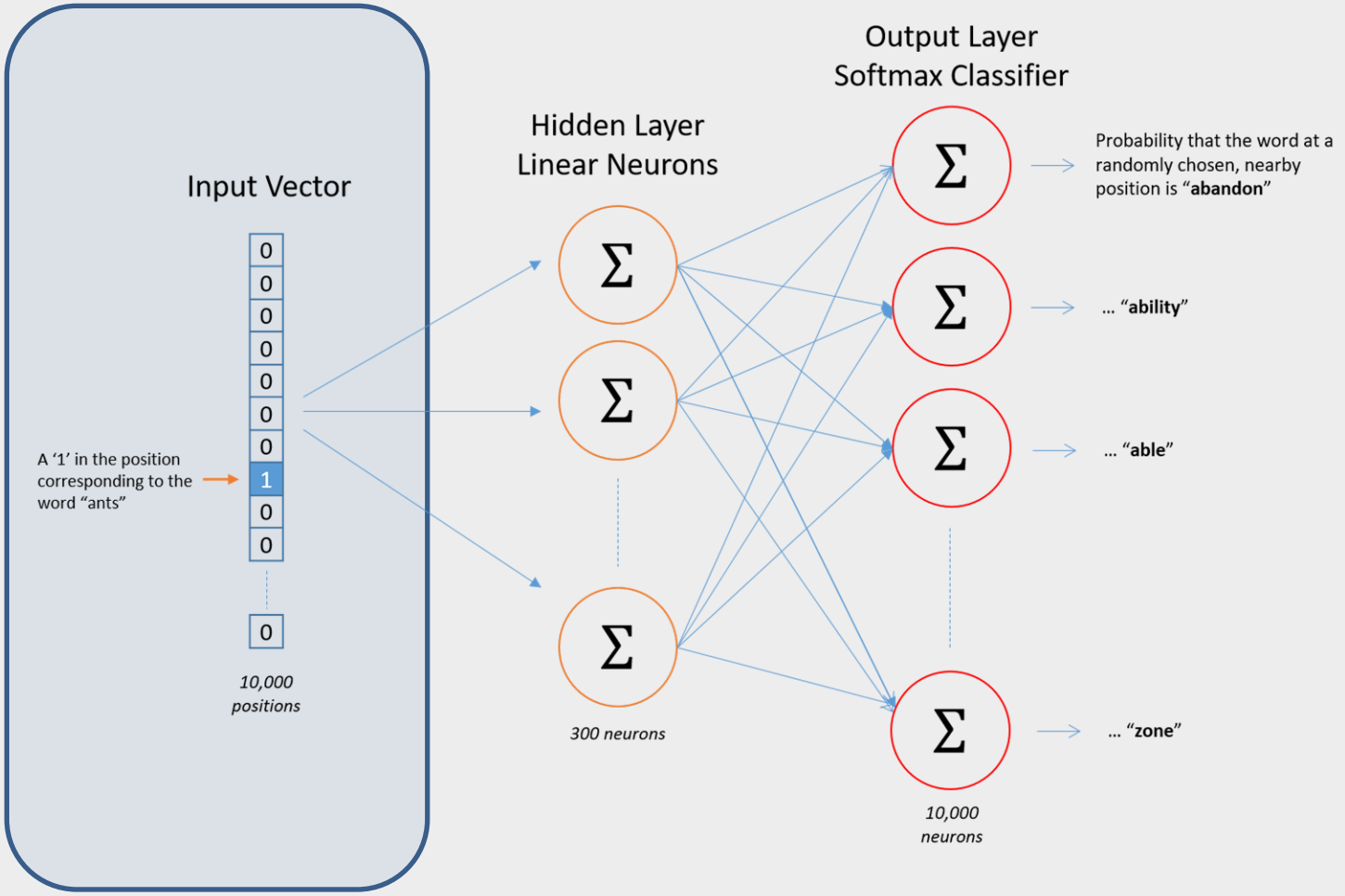
- Ejemplo: el modelo skip-gram
  - Objetivo: dada una palabra, indicar la probabilidad de que otra aparezca en su contexto



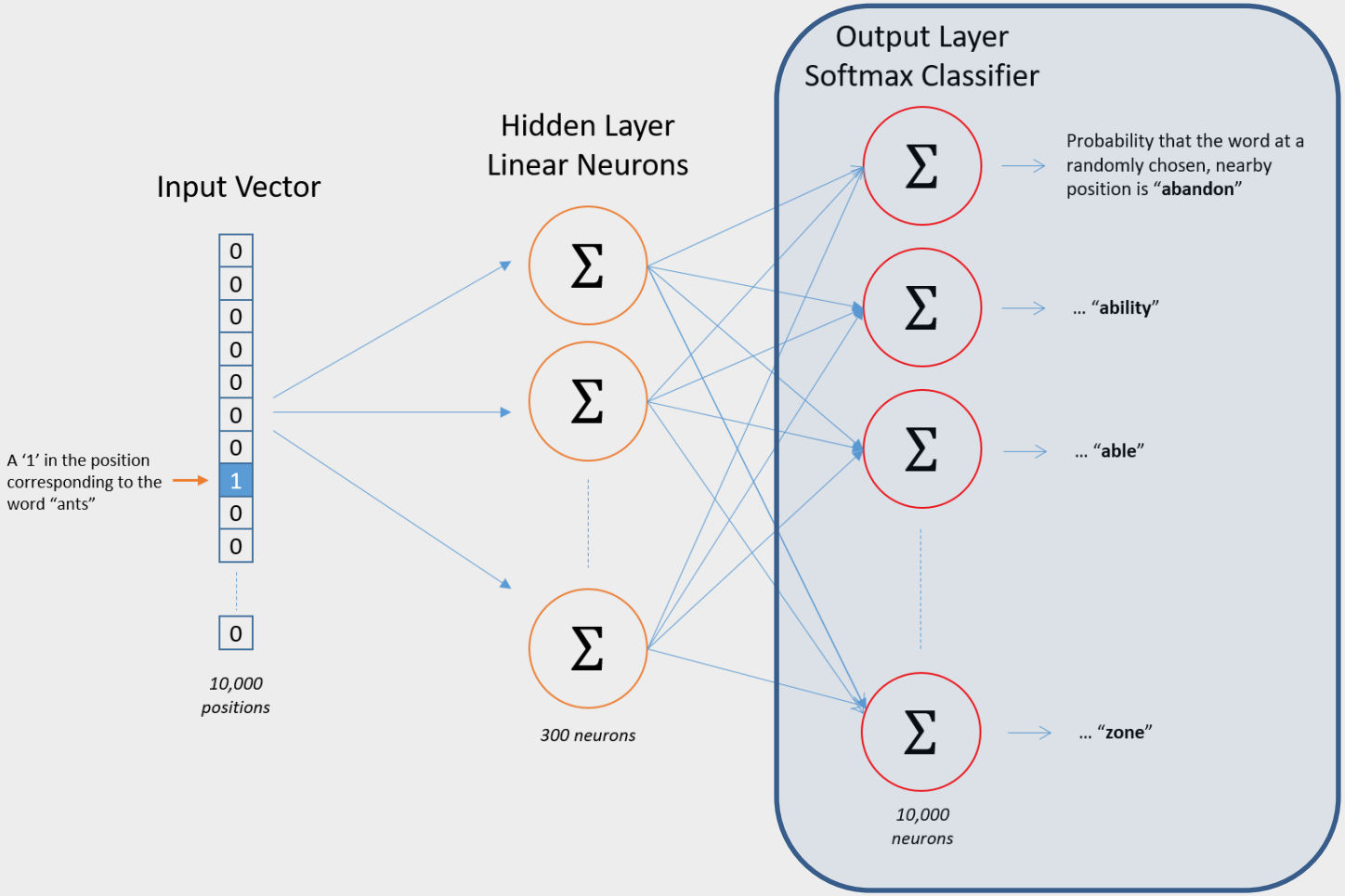
- Contexto: tamaño de ventana 2



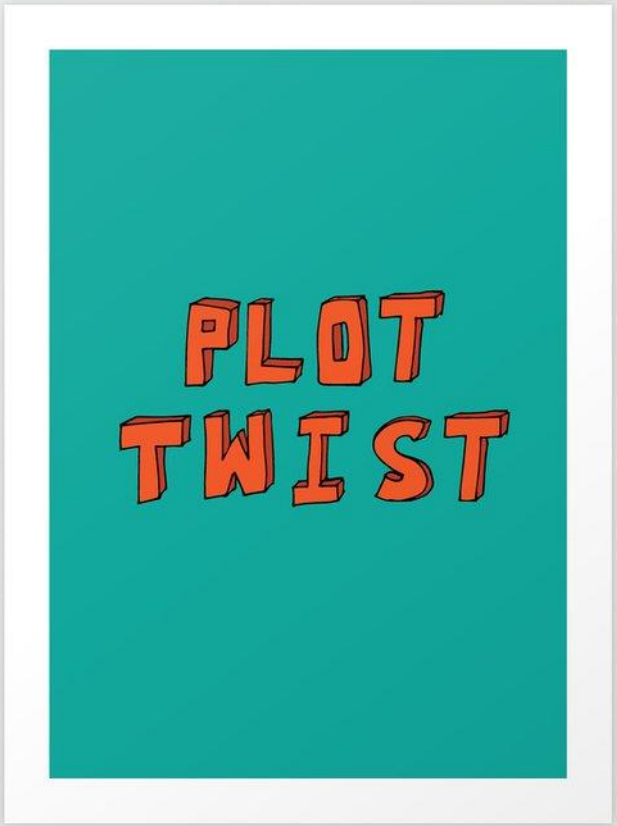
- Representación de la información
  - Entrada es un vector de n dimensiones (una por cada palabra) - One-hot vector

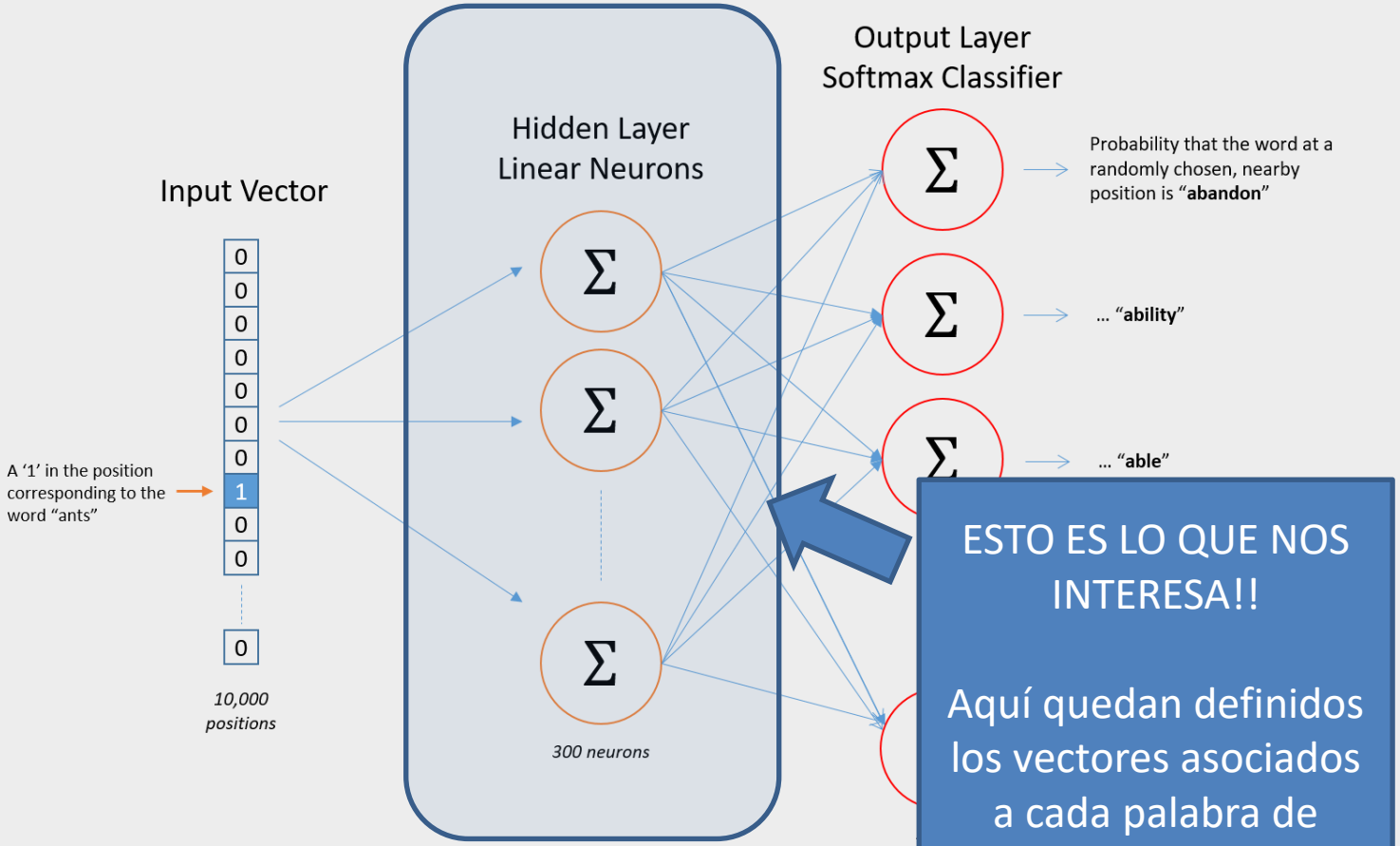

Input Vector
0
0
0
0
0
0
0
1
0
0
0
A '1' in the position corresponding to the word "ants"
10,000 positions
Hidden Layer Linear Neurons
300 neurons
Output Layer Softmax Classifier
10,000 neurons
Probability that the word at a randomly chosen, nearby position is "abandon"
... "ability"
... "able"
... "zone"

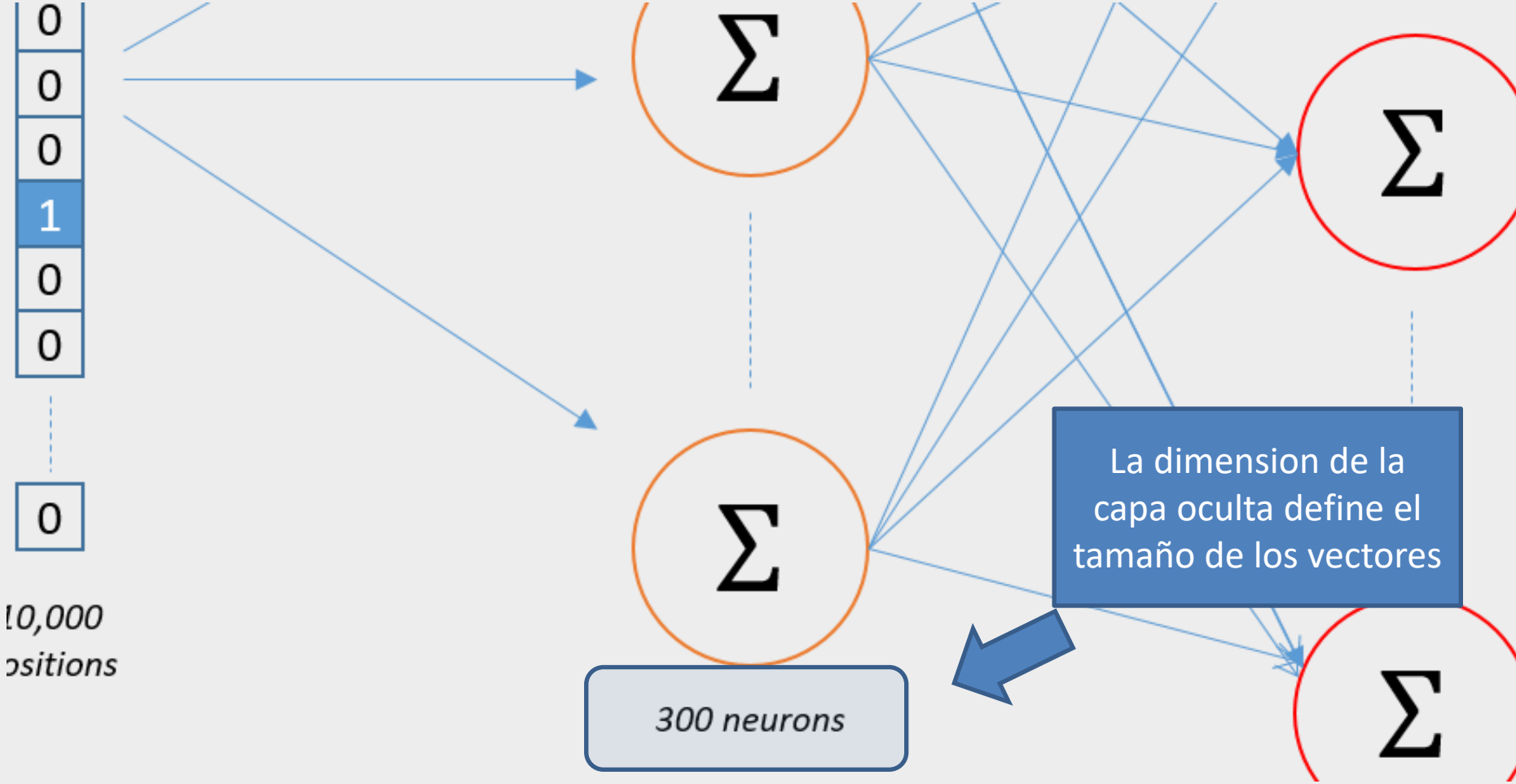
- Representación de la información
  - Entrada es un vector de *n* dimensiones (una por cada palabra) - One-hot vector
  - Salida es un vector de *n* dimensiones (cada posición es la probabilidad de que esa palabra esté en la cercanía de la palabra de la entrada)


Input Vector
Hidden Layer
Linear Neurons
Output Layer
Softmax Classifier
A '1' in the position corresponding to the word "ants"
0
0
0
0
0
0
0
1
0
0
0
10,000 positions
300 neurons
Probability that the word at a randomly chosen, nearby position is "abandon"
... "ability"
... "able"
... "zone"
10,000 neurons

Entrenamos la red, pero la tiramos a la basura…




Input Vector
A '1' in the position corresponding to the word "ants"
0
0
0
0
0
0
0
1
0
0
0
10,000 positions
Hidden Layer
Linear Neurons
Σ
Σ
Σ
300 neurons
Output Layer
Softmax Classifier
Σ
Σ
Σ
Probability that the word at a randomly chosen, nearby position is "abandon"
... "ability"
... "able"
ESTO ES LO QUE NOS INTERESA!!
Aquí quedan definidos los vectores asociados a cada palabra de nuestro vocabulario


0
0
0
1
0
0
0
10,000
positions
Σ
Σ
Σ
Σ
La dimension de la capa oculta define el tamaño de los vectores
300 neurons

Si dos palabras se usan en contextos similares, entonces la salida de la red debería ser similar…

Si dos palabras se usan en contextos similares, entonces la salida de la red debería ser similar…

… y sus vectores asociados

(word vectors) también

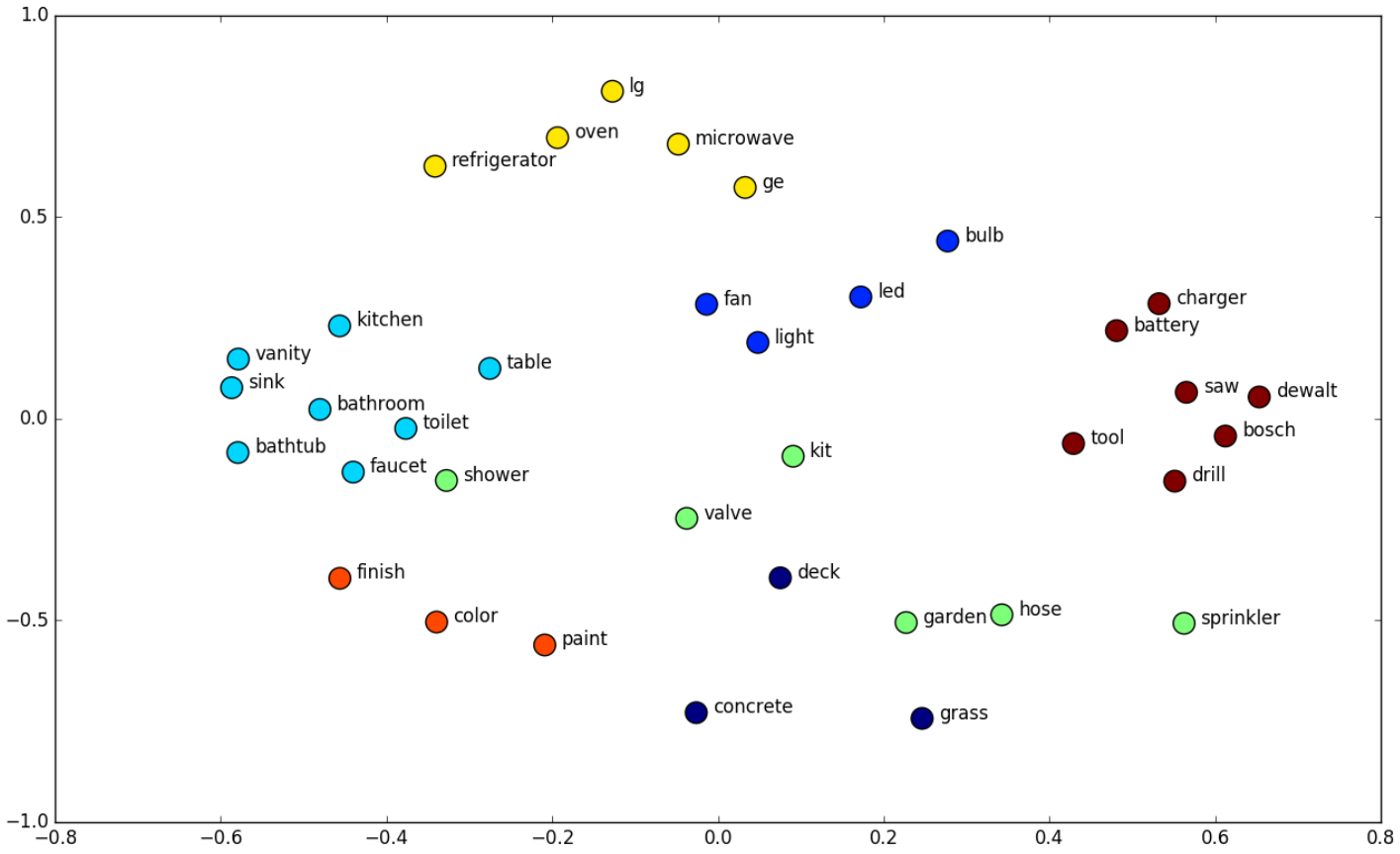
- Algunas consideraciones
  - Si bien el entrenamiento corresponde a un problema supervisado, no exige ningún tipo de proceso de etiquetado
  - Ídem con las relaciones semánticas (son capturadas por los word vectors a partir de los contextos)
  - El entrenamiento se realiza utilizando cualquier corpus disponible: Wikipedia, Google News, un dominio especifico, etc.

- Ejemplo:

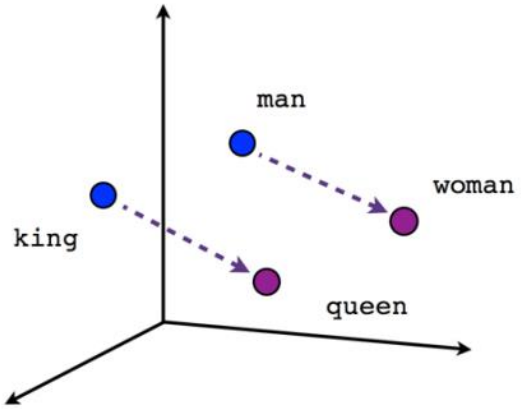
```
# look up top 6 words similar to 'shocked'
w1 = ["shocked"]
model.wv.most_similar (positive=w1,topn=6)
```

```
[('horrified', 0.807753860950 47),
 ('amazed', 0.779747009277 3438),
 ('astonished', 0.774845957756 0425),
 ('dismayed', 0.768063306808 4717),
 ('stunned', 0.760303497314 4531),
 ('appalled', 0.746677637100 2197)]
```

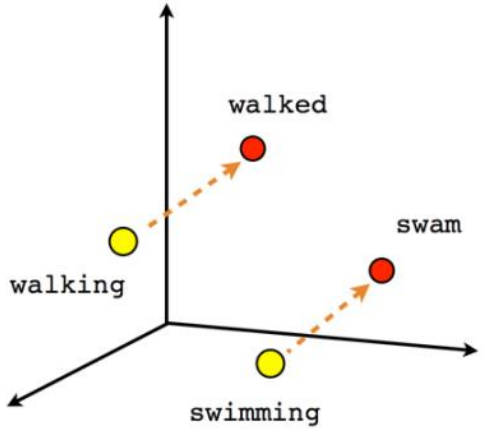
- Preservación de la relación semántica



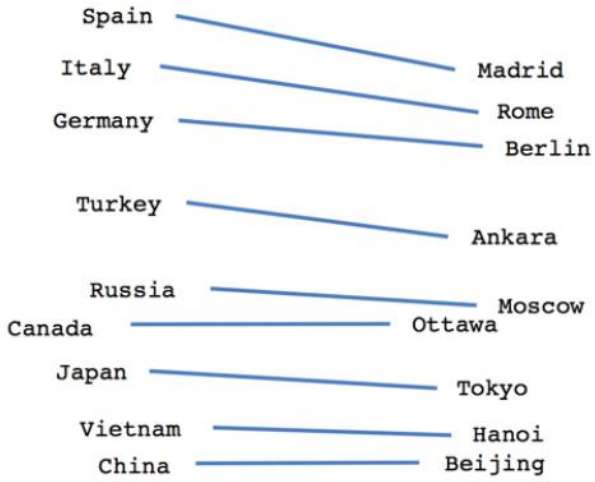
- Preservación de la relación semántica



Male-Female

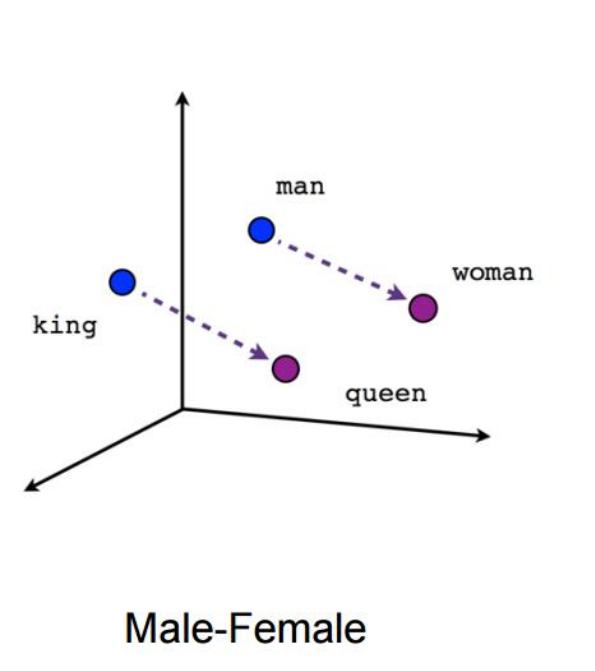


Verb tense



Country-Capital

- Preservación de la relación semántica
  - Word analogy

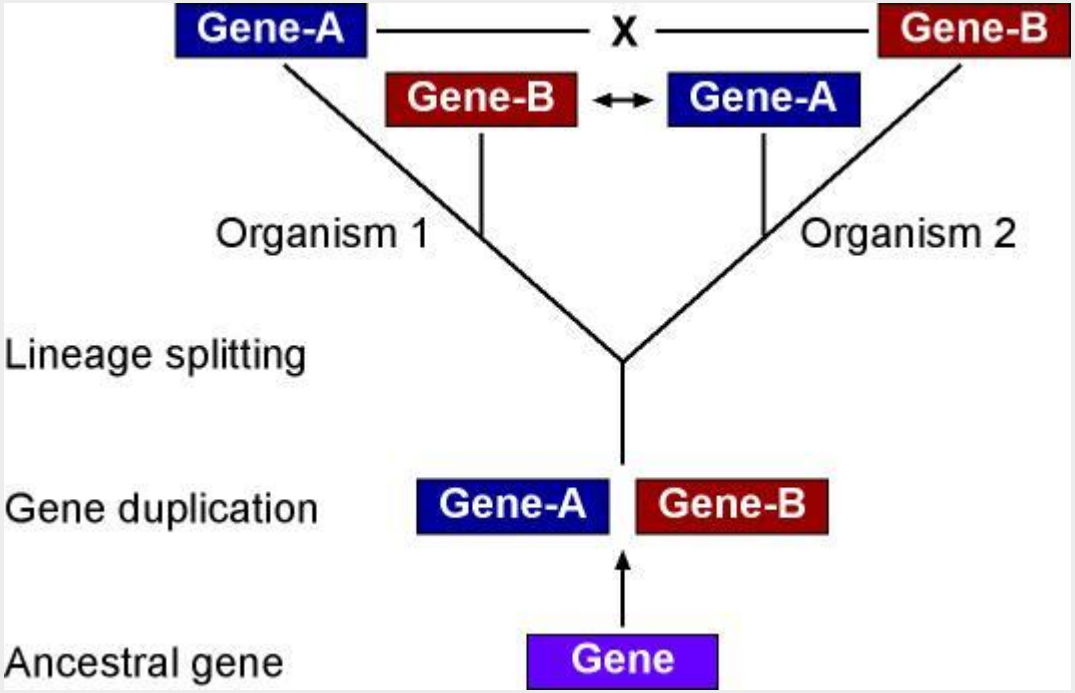


king–man+woman -> ?

Algunas aplicaciones

- Knowledge discovery: Bioinformática
  - Costos de secuenciación cada vez más económicos
  - Estudios sobre nuevos organismos
  - La información es reportada en forma de texto libre (papers)
  - No existen estándares para la sistematización de la información

- Ejemplo: genes ortólogos



GenA-Organismo1+Organismo2 = ?

GenX = ?

- Knowledge discovery: ciencia de materiales

nature

Letter | Published: 03 July 2019

Unsupervised word embeddings capture latent knowledge from materials science literature

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder & Anubhav Jain

Nature **571**, 95–98(2019) | Cite this article

42k Accesses | **13** Citations | **1562** Altmetric | Metrics

Abstract

The overwhelming majority of scientific knowledge is published as text,

ferromagnetic–NiFe+IrMn

⇩

antiferromagnetic

- Algunas consideraciones
  - “Entrenamiento” fuertemente dependiente del corpus disponible
  - “dificultad” de acceso a la información (copyright)
  - Sesgo de los datos: estereotipos
    - doctor–man+woman = nurse

– Otras líneas de investigación de nuestra facultad
  - Procesamiento de señales / reconocimiento de imágenes y patrones (Alvaro Pardo)
  - Sistemas de recomendación clínica y NLP para procesamiento de historias clínicas (Ernesto Ocampo)
  - Machine Learning en bioinformática – Urugenomes (Lucía Spangenberg)
  - Cheminformatics (Gustavo Vazquez)

Contactos

¡ Muchas Gracias !

http://www.ucu.edu.uy/

Gustavo Vazquez
gustavo.vazquez@ucu.edu.uy