



Representación Semántica de Frases

Héctor Cerezo-Costas (hcerezo@gradient.org)
WGML2016

Perfil Personal

Experiencia

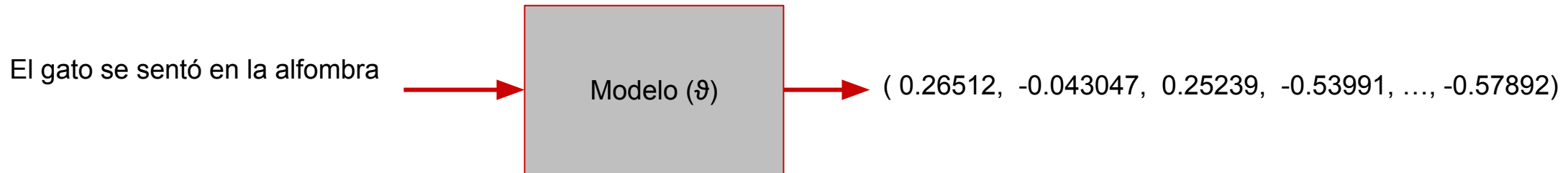
- Grupo de investigación GTI (Uvigo) (2007)
- Gradiant (2010)
 - Desde 2012 trabajando en PLN
 - Desde 2015 trabajando en deep learning
- Realizando la tesis doctoral en deep learning y PLN

Y además...

- Interesado en Inteligencia artificial

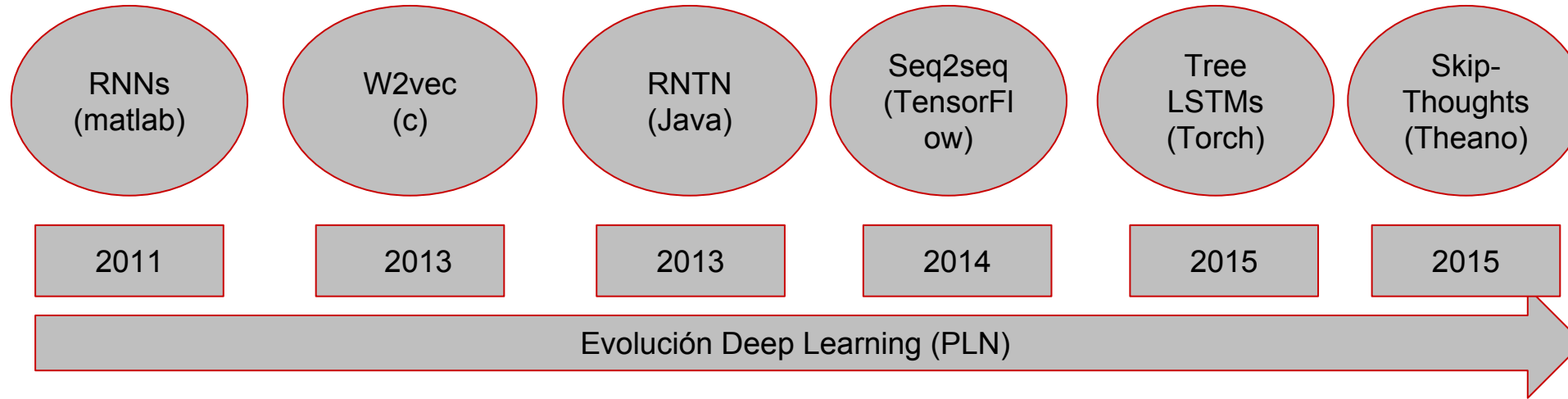
Motivación

Métodos para la extracción de vectores semánticos (embeddings de frases):



- Entrenamiento limitado $\Rightarrow$ modelo específico (baja cobertura).
- Dificultad para obtener entrenamiento.

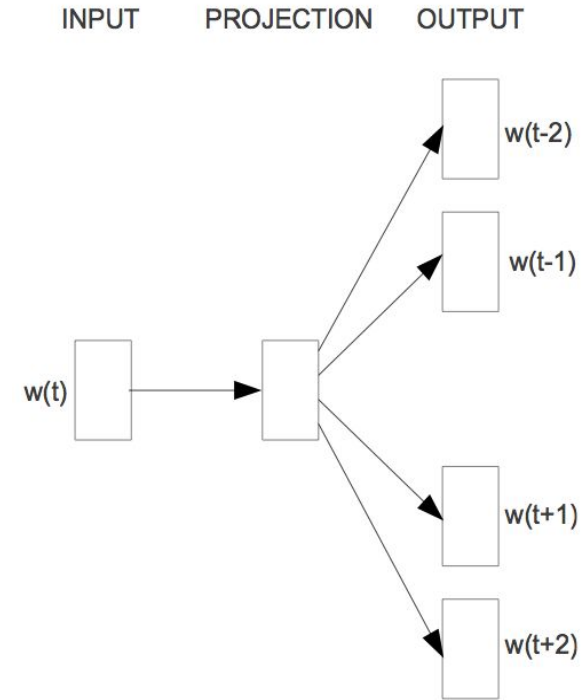
Deep Learning en PLN



Vectores de Palabras (W2vec - Modelo Skipgram)

Intuición: palabras freq. juntas $\Rightarrow$ existe relación semántica

Vectores de palabras son mejores
 $\Rightarrow$ predicción del contexto más precisa.

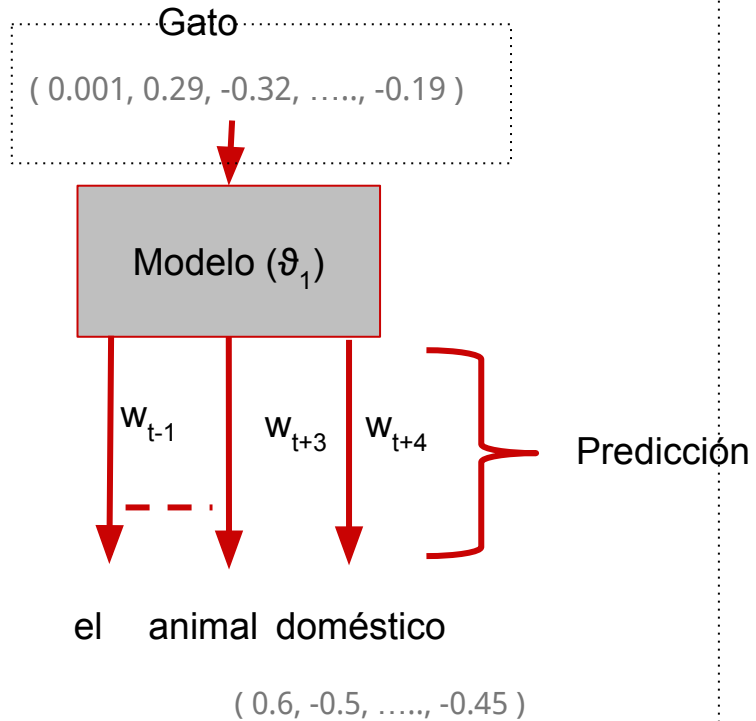


$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t)$$

W2vec - Explicación

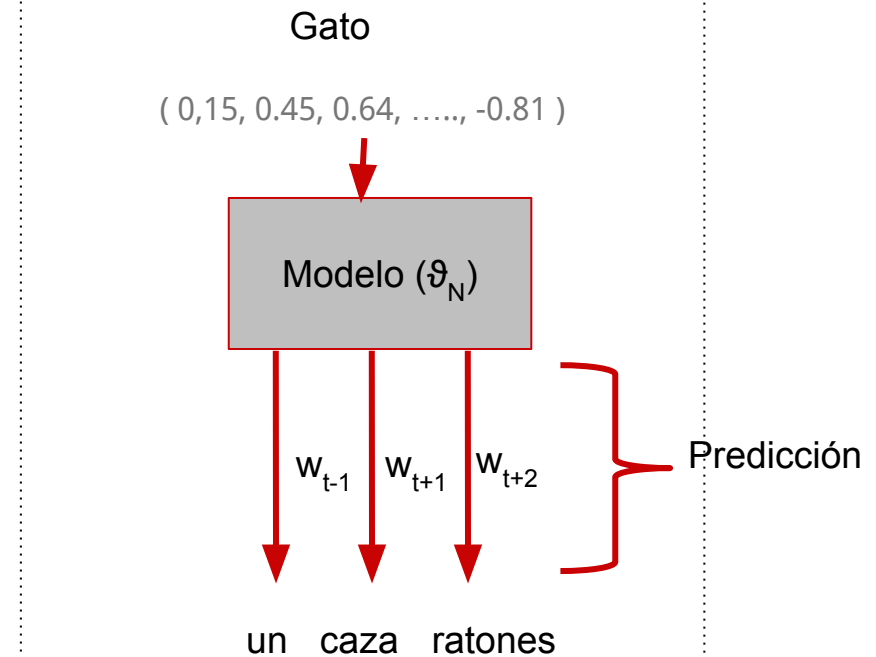
Inicialmente: Vectores Aleatorios

El **gato** es un animal doméstico



N batches

Un **gato** caza ratones cuando tiene hambre

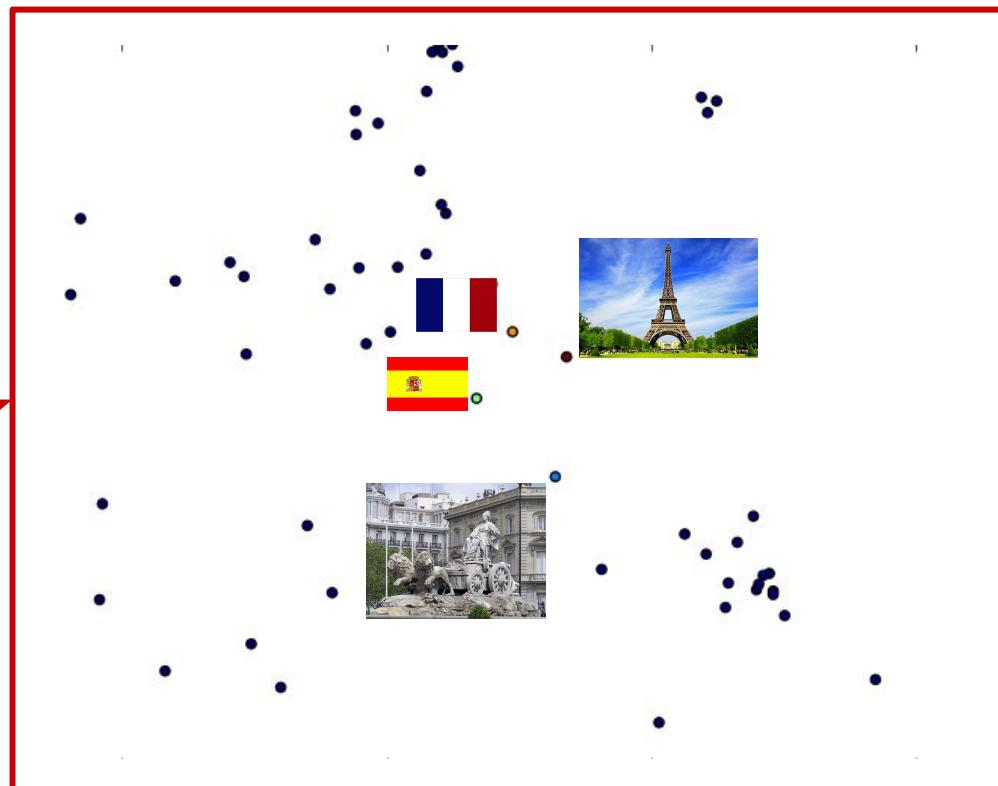
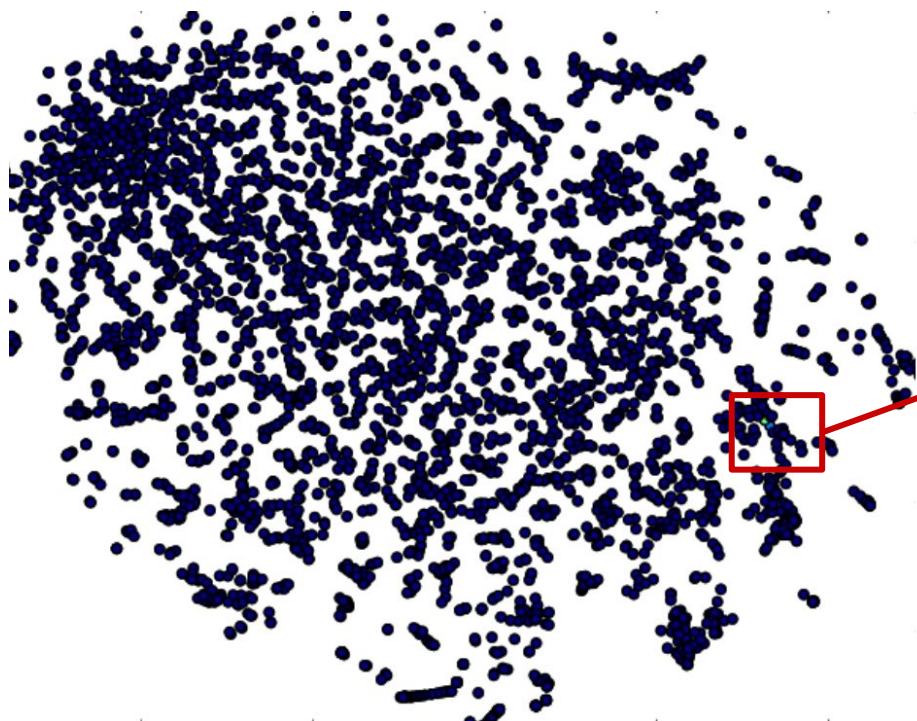


Validación - Analogías de palabras

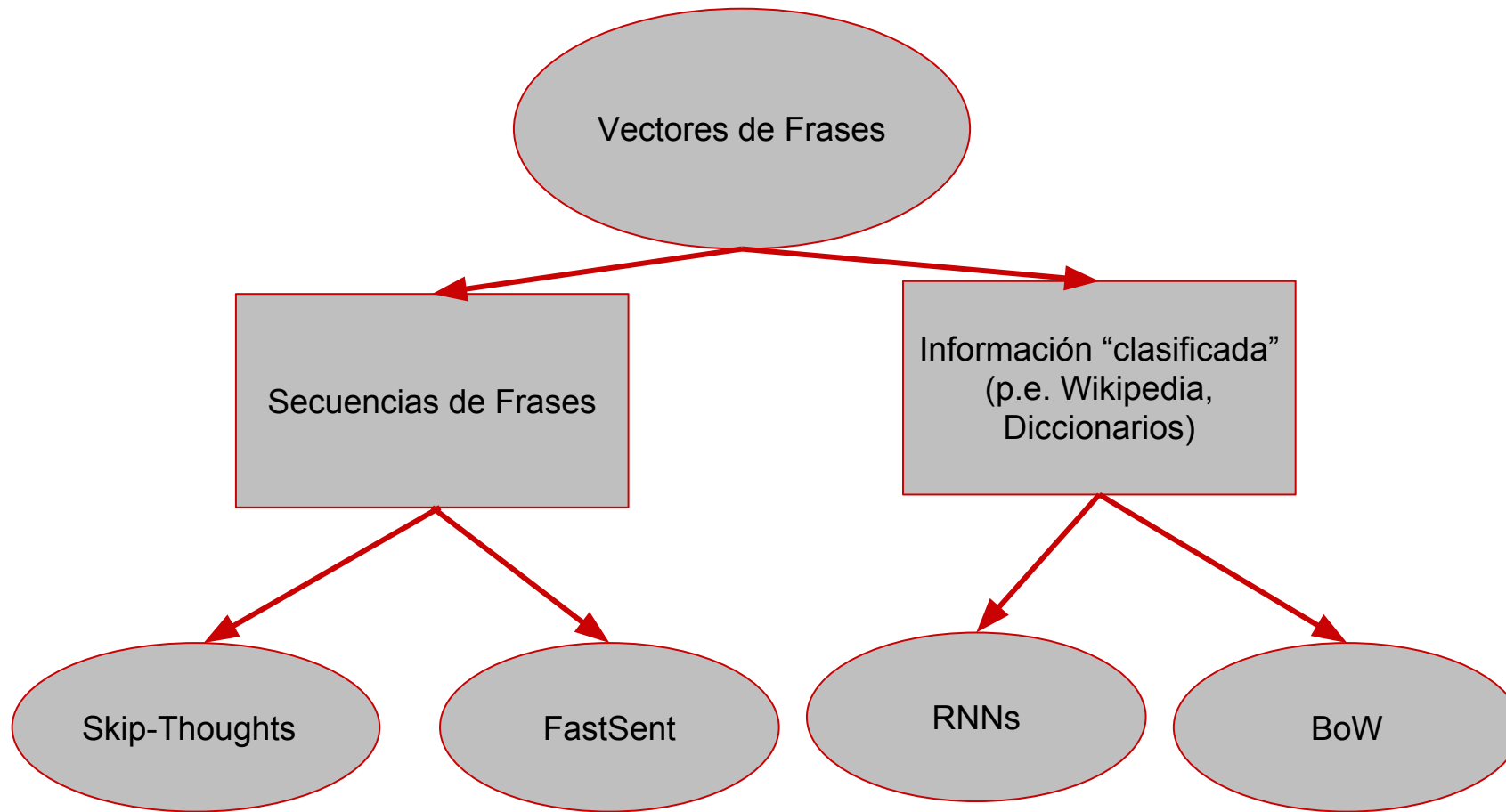
París es a Francia lo que Madrid es a ?



$v[\text{Francia}] - v[\text{París}] + v[\text{Madrid}] \Rightarrow ?$

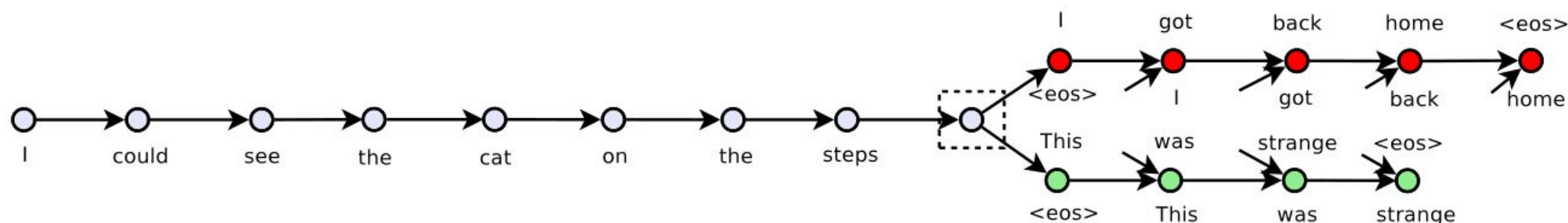


Vectores de Frases



Abstracción de frases: Skip-Thoughts

- Funcionamiento: codifica una frase de forma que maximiza la probabilidad de recuperar las frases contiguas.
- Entrenamiento:
 - Frases consecutivas de libros, webs, etc.
- No supervisado

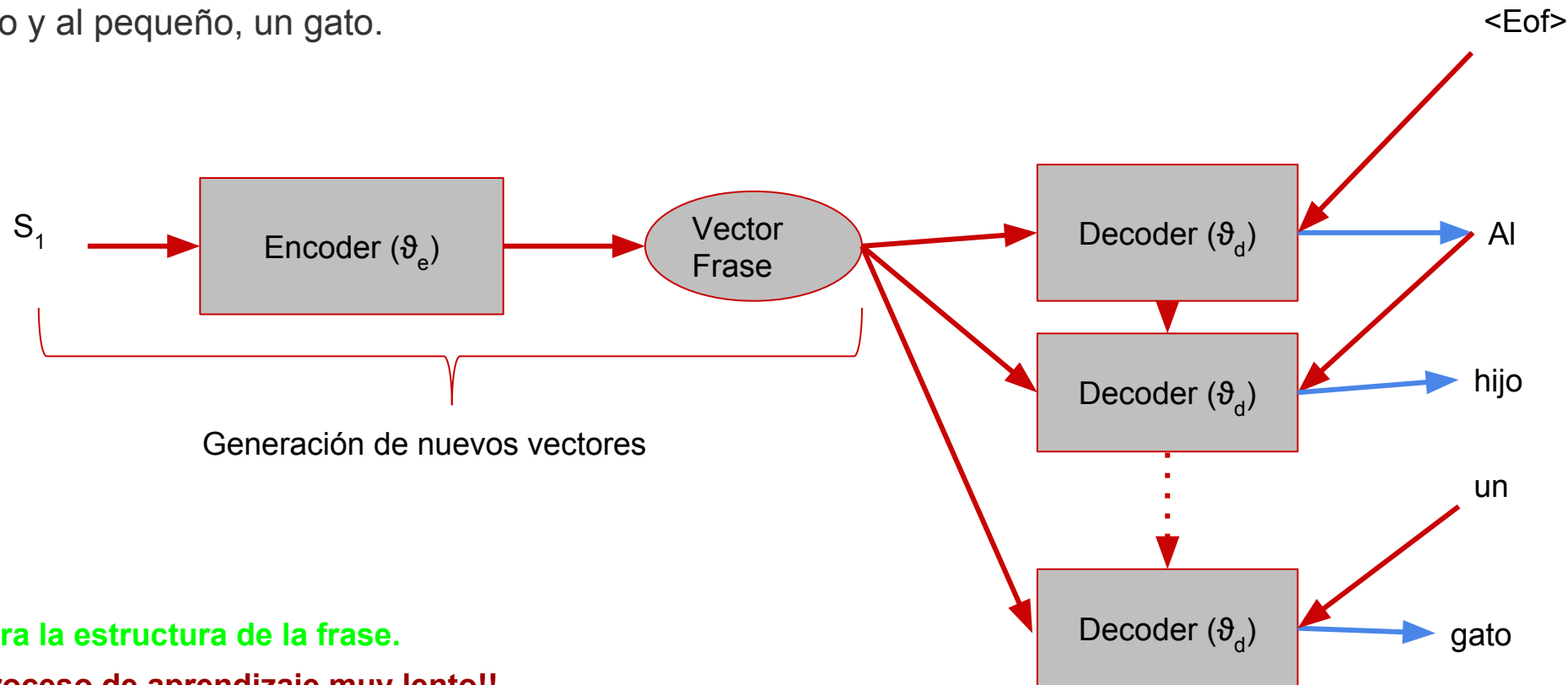


Skip-Thoughts (Funcionamiento)

S0: Érase una vez un molinero que tenía tres hijos.

S1: El hombre era muy pobre y casi no tenía bienes para dejarles en herencia.

S2: Al hijo mayor le legó su viejo molino, al mediano un asno y al pequeño, un gato.



Considera la estructura de la frase.

Es un proceso de aprendizaje muy lento!!

Diccionarios Inversos

Tenemos muchos diccionarios online:

- Entrenamiento
 - Descripción - Palabra
- Semi-supervisado
- Tiempos de aprendizaje más cortos.

trompa

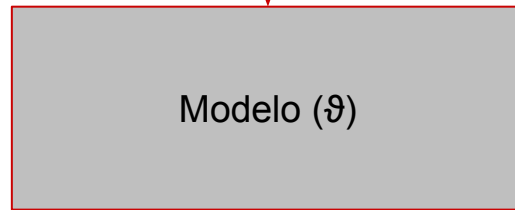
De or. onomat.

1. f. Instrumento musical de viento, consistente en un tubo de latón enroscado circularmente que va ensanchándose desde la boquilla al pabellón, cuyo sonido se modifica por medio de pistones y antiguamente introduciendo la mano en el pabellón.
2. f. Prolongación muscular, hueca y elástica, de la nariz de algunos animales, como el elefante o el tapir, capaz de absorber fluidos.
3. f. Trompo grande que tiene dentro otros pequeños, los cuales, saliendo de él impetuosamente al tiempo de ser arrojado para que baile, giran todos a un tiempo.
4. f. Trompo grande, hueco, con una abertura lateral para que zumbe.
5. f. Trompo, especialmente el grande y de forma achatada.
6. f. Aparato para soplar en una forja catalana, que consistía en un tubo vertical por donde se dejaba caer un chorro de agua que impelía el aire necesario.
7. f. Bohordo de la cebolla cortado, en que solían soplar los muchachos para hacerlo

[Fuente: RAE](#)

Diccionarios Inversos (Modelo)

Gran mamífero carnívoro férido africano,
cuyo macho se distingue por una larga
melena

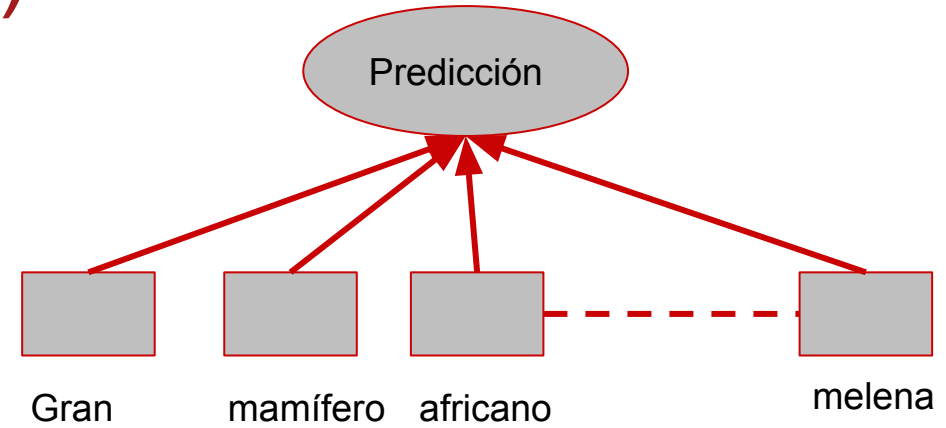


Error:

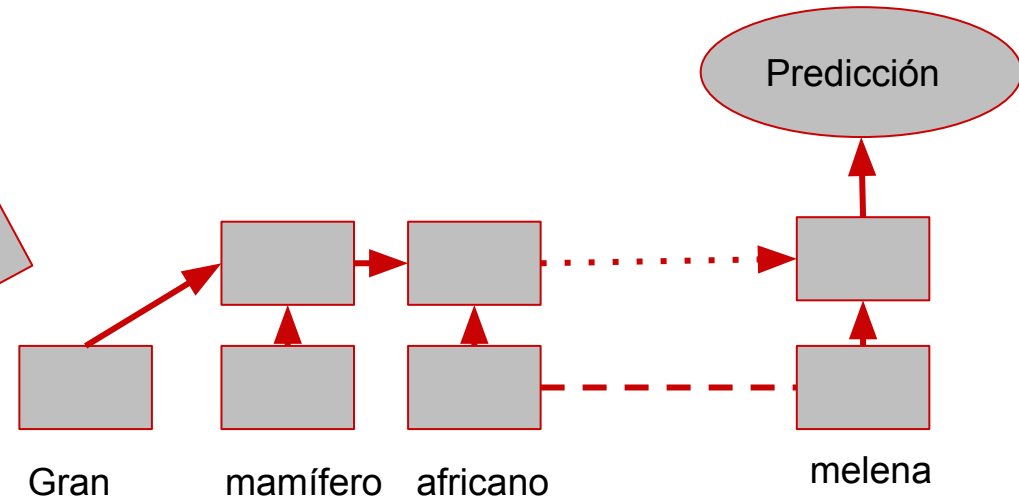
- cosine similarity
- Cosine rankings...



Sin estructura



Con estructura

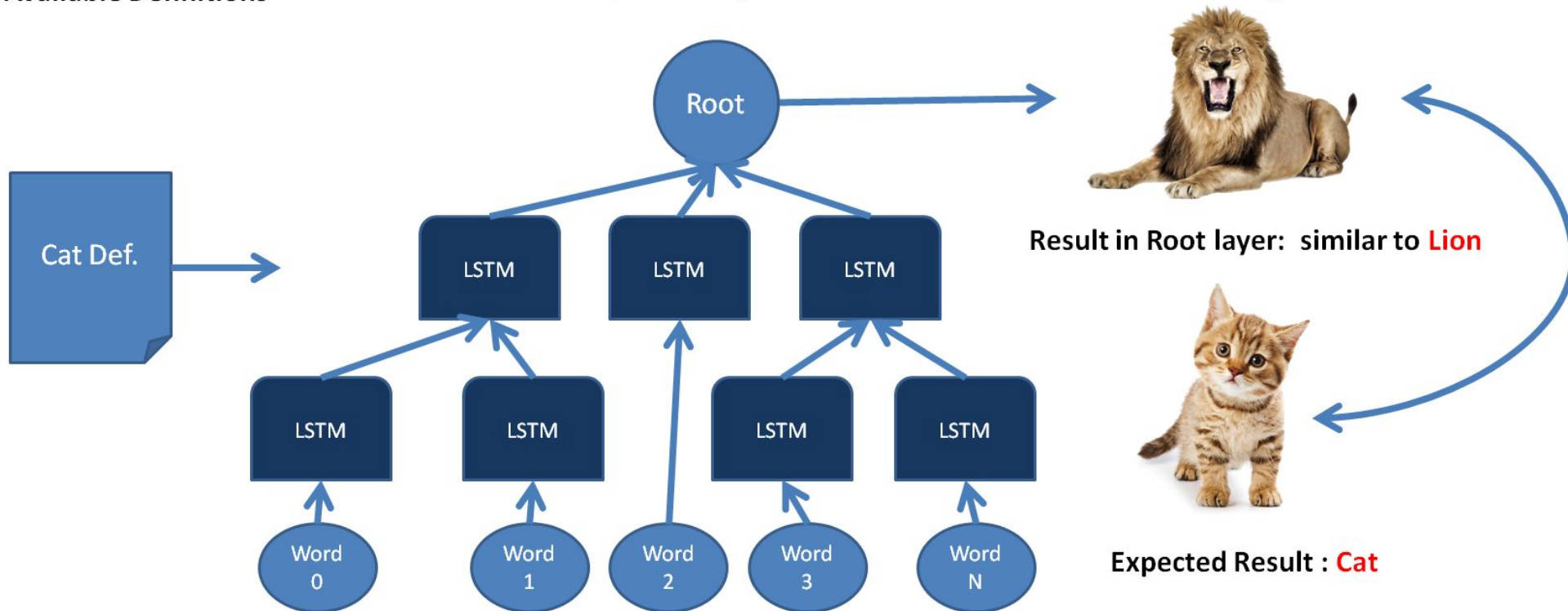


Diccionarios Inversos (Modelos II)

Available Definitions

Multilayered Deep Neural Net

Minimising Cosine Error



Medir el rendimiento

- Supervisada (clasificación): sentimiento, temática, etc.
- No supervisada : similitud entre frases

Frase 1	Frase 2	Score
A brown dog is attacking another animal in front of the tall man in pants	A brown dog is attacking an other animal in front of the man in pants	4.9
The player is missing the basket and a crowd is in background	The player is dunking the basketball into the net and a crowd is in background	3.9
The hungry woman is eating	A woman is mixing eggs	2.5
A lone biker is jumping in the air	A man is jumping into a full pool	1.7

[Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual Semantic Textual Similarity. In Proceedings of the 8th international workshop on semantic evaluation \(SemEval 2014\), pages 81–91.](#)

[Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In LREC, pages 216–223.](#)

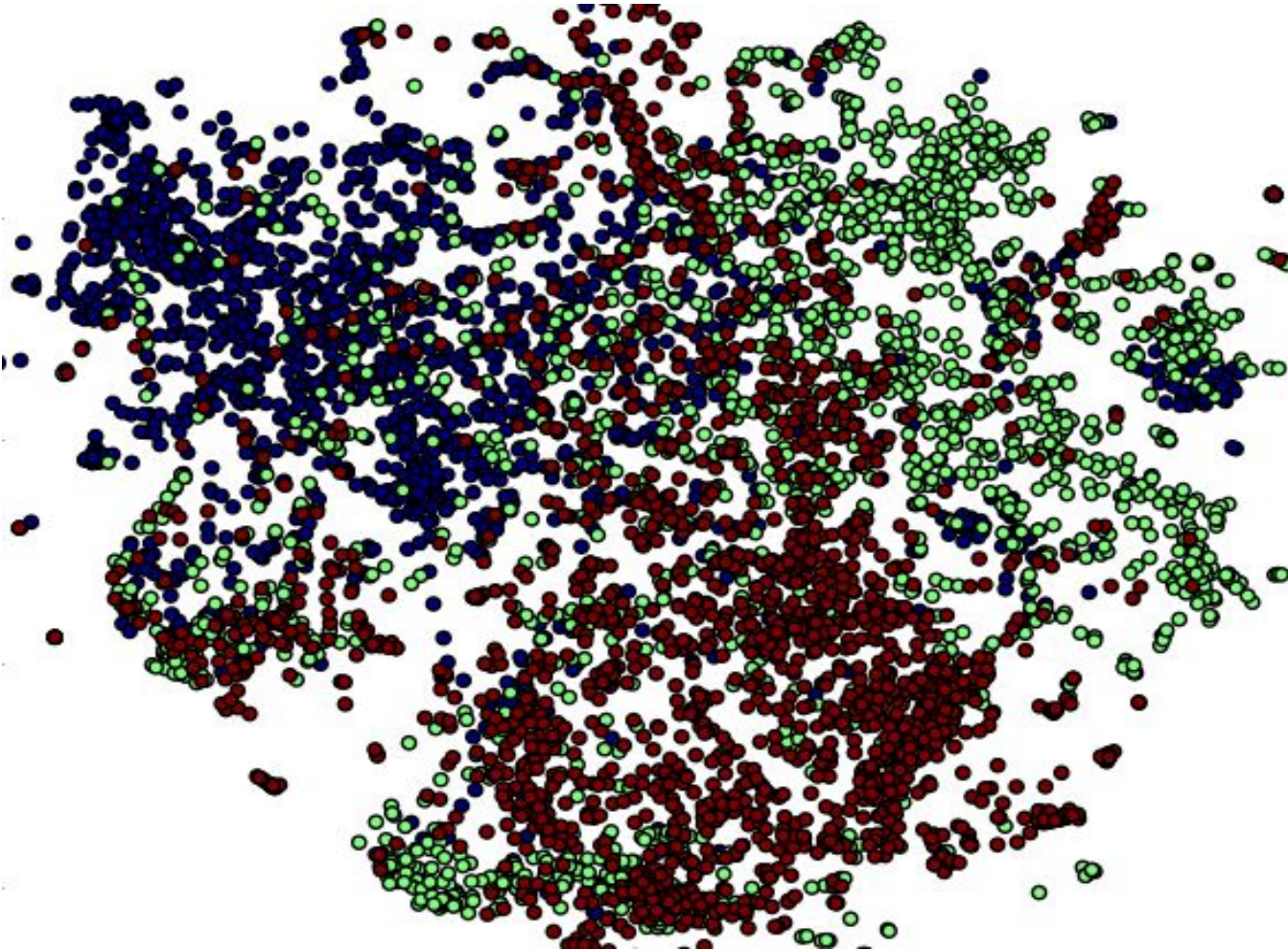
Comparación (Rendimiento)

Model	STS 2014							SICK
	News	Forum	WordNet	Twitter	Images	Headlines	All	Test + Train
SAE	.17/.16	.12/.12	.30/.23	.28/.22	.49/.46	.13/.11	.12/.13	.32/.31
SAE+embs.	.52/.54	.22/.23	.60/.55	.60/.60	.64/.64	.41/.41	.42/.43	.47/.49
SDAE	.07/.04	.11/.13	.33/.24	.44/.42	.44/.38	.36/.36	.17/.15	.46/.46
SDAE+embs.	.51/.54	.29/.29	.56/.50	.57/.58	.59/.59	.43/.44	.37/.38	.46/.46
ParagraphVec DBOW	.31/.34	.32/.32	.53/.5	.43/.46	.46/.44	.39/.41	.42/.43	.42/.46
ParagraphVec DM	.42/.46	.33/.34	.51/.48	.54/.57	.32/.30	.46/.47	.44/.44	.44/.46
Skipgram	.56/.59	.42/.42	.73/.70	.71/.74	.65/.67	.55/.58	.62/.63	.60/.69
CBOW	.57/.61	.43/.44	.72/.69	.71/.75	.71/.73	.55/.59	.64/.65	.60/.69
Unigram TFIDF	.48/.48	.40/.38	.60/.59	.63/.65	.72/.74	.49/.49	.58/.57	.52/.58
SkipThought	.44/.45	.14/.15	.39/.34	.42/.43	.55/.60	.43/.44	.27/.29	.57/.60
FastSent	.58/.59	.41/.36	.74/.70	.63/.66	.74/.78	.57/.59	.63/.64	.61/.72
FastSent+AE	.56/.59	.41/.40	.69/.64	.70/.74	.63/.65	.58/.60	.62/.62	.60/.65
NMT En to Fr	.35/.32	.18/.18	.47/.43	.55/.53	.44/.45	.43/.43	.43/.42	.47/.49
NMT En to De	.47/.43	.26/.25	.34/.31	.49/.45	.44/.43	.38/.37	.40/.38	.46/.46
CaptionRep BOW	.26/.26	.29/.22	.50/.35	.37/.31	.78/.81	.39/.36	.46/.42	.56/.65
CaptionRep RNN	.05/.05	.13/.09	.40/.33	.36/.30	.76/.82	.30/.28	.39/.36	.53/.62
DictRep BOW	.62/.67	.42/.40	.81/.81	.62/.66	.66/.68	.53/.58	.62/.65	.57/.66
DictRep BOW+embs.	.65/.72	.49/.47	.85/.86	.67/.72	.71/.74	.57/.61	.67/.70	.61/.70
DictRep RNN	.40/.46	.26/.23	.78/.78	.42/.42	.56/.56	.38/.40	.49/.50	.49/.56
DictRep RNN+embs.	.51/.60	.29/.27	.80/.81	.44/.47	.65/.70	.42/.46	.54/.57	.49/.59
CPHRASE	.69/.71	.43/.41	.76/.73	.60/.65	.75/.79	.60/.65	.65/.67	.60/.72

Mejor sin estructura
!!!

Table 4: Performance of sentence representation models (Spearman/Pearson correlations) on **unsupervised** (relatedness) evaluations (Section 3.2). Models are grouped according to training data as indicated in Table 3

T-SNE Frases



Visualización T-sne de vectores de frases para tres categorías de Stackoverflow

Comparativa (computación)

Modelo	Entrada	Tiempo (Entrenamiento)	Dimensión
Skip-Thoughts	11.038 libros	2 semanas (GPU)	4800
FastSent	11.038 libros	(No especificado)	100
DictRep BoW	≈800 mil defs.	unas horas (GPU)/< 1 día (CPU)	500
DictRep RNN	≈800 mil defs.	1 día (GPU)/3-5 días CPU	500
DictRep (grafo)	≈800 mil defs.	3-5 semanas (CPU)	500
word2vec	100 mil millones palabras	2 semanas (CPU)	300

Problemas a resolver

- Mejores métodos de entrenamiento
- Mejorar formas de medir el rendimiento de los distintos modelos
 - modelos sin estructura “parecen” funcionar mejor en tests no supervisados
- Implementaciones más eficientes
 - menor uso de recursos
 - más rápidos

Esto es todo... de momento

