



Modelos de Lenguaje

- N-grams
- Neurales

Modelos de Lenguaje



¿Cuál es la probabilidad de "gato" en:

Cuando llegue a casa me tengo que acordar de darle de comer al _____

Modelos de Lenguaje



¿Cuál es la probabilidad de "gato" en:

Cuando llegue a casa me tengo que acordar de darle de comer al _____

¿Es más probable "gato" o "perro"?

Modelos de Lenguaje



¿Cuál es la probabilidad de "gato" en:

Cuando llegue a casa me tengo que acordar de darle de comer al _____

¿Es más probable "gato" o "perro"?

Aplicaciones:

- Traducción
- Reconocimiento de habla
- Corrección en la escritura
- Sistema de ayuda

Modelos de Lenguaje



¿Podemos estimar la probabilidad a posteriori de una palabra dada su historia?

$p(w | h)$ = probabilidad de palabra w dada su historia h

N-grams

N-grams



Podemos calcularlo como:

$$p(w | h) = \#(h, w) / \#(h)$$

$$p(\text{"gato"} | \text{"Cuando llegue a casa me tengo que acordar de darle de comer al"}) = \frac{\#(\text{"Cuando llegue a casa me tengo que acordar de darle de comer al gato"})}{\#(\text{"Cuando llegue a casa me tengo que acordar de darle de comer al"})}$$

N-grams



Podemos calcularlo como:

$$p(w | h) = \#(h, w) / \#(h)$$

$$p(\text{"gato"} | \text{"Cuando llegue a casa me tengo que acordar de darle de comer al"}) = \frac{\#(\text{"Cuando llegue a casa me tengo que acordar de darle de comer al gato"})}{\#(\text{"Cuando llegue a casa me tengo que acordar de darle de comer al"})}$$

Pero cómo cuento cuántas veces aparece una frase en **el lenguaje**.

N-grams



Si lo calculamos con regla de la cadena:

$$p(w_1 w_2 w_3 \dots w_n) = p(w_1) * p(w_2 | w_1) * p(w_3 | w_1 w_2) * p(w_4 | w_1 w_2 w_3) * \dots * p(w_n | w_1 w_2 w_3 \dots w_{n-1})$$

$$p(\text{"Cuando llegue a casa me tengo que acordar de darle de comer al gato"}) = \\ p(\text{cuando}) * p(\text{llegue} | \text{Cuando}) * p(\text{a} | \text{Cuando llegue}) * \dots * p(\text{gato} | \text{Cuando llegue} \dots \text{a} \dots)$$

$$p(w_1^n) = \prod_{k=1}^N p(w_k | w_1^{k-1})$$

N-grams



Si lo calculamos con regla de la cadena:

$$p(w_1 w_2 w_3 \dots w_n) = p(w_1) * p(w_2 | w_1) * p(w_3 | w_1 w_2) * p(w_4 | w_1 w_2 w_3) * \dots * p(w_n | w_1 w_2 w_3 \dots w_{n-1})$$

$$p(\text{"Cuando llegue a casa me tengo que acordar de darle de comer al gato"}) = \\ p(\text{cuando}) * p(\text{llegue} | \text{Cuando}) * p(\text{a} | \text{Cuando llegue}) * \dots * p(\text{gato} | \text{Cuando llegue} \dots \text{a} \dots)$$

$$p(w_1^n) = \prod_{k=1}^N p(w_k | w_1^{k-1})$$

Problema: el lenguaje es creativo, es probable que (h, w) no exista

N-grams

Si lo calculamos con regla de la cadena:

$$p(w_1 w_2 w_3 \dots w_n) = p(w_1) * p(w_2 | w_1) * p(w_3 | w_1 w_2) * p(w_4 | w_1 w_2 w_3) * \dots * p(w_n | w_1 w_2 w_3 \dots w_{n-1})$$

$$p(\text{"Cuando llegue a casa me tengo que acordar de darle de comer al gato"}) = \\ p(\text{cuando}) * p(\text{llegue} | \text{Cuando}) * p(\text{a} | \text{Cuando llegue}) * \dots * p(\text{gato} | \text{Cuando llegue a ...})$$

$$p(w_1^n) = \prod_{k=1}^N p(w_k | w_1^{k-1})$$

Problema: el lenguaje es creativo, es probable que (h, w) no exista

Ventaja: podemos buscar n más chico que permita calcular las probabilidades y aproximar la probabilidad

$$p(\text{gato} | \text{Cuando llegue a casa me tengo que acordar de darle de comer al}) \sim p(\text{gato} | \text{darle de comer al})$$

$$p(w_1^n) \sim \prod_{k=1}^N p(w_k | w_{n-N+1}^{k-1})$$

5-gram

N-grams



Estimador de Máxima Verosimilitud (MLE):

$$p(w_1^n) \sim \prod_{k=1}^N p(w_k \mid w_{n-N+1}^{k-1}) \longrightarrow P(w_n \mid w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1} w_n)}{C(w_{n-N+1}^{n-1})}$$

Evaluación de LM



Evaluación extrínseca: Usar los resultados de nuestro n-gram en alguna aplicación externa

- Entrenar modelos para después testarlos en otra aplicación puede ser costoso

Evaluación intrínseca: Evaluar el modelo en un test set

- Es importante que las oraciones de test no estén en el training set
- Para evaluar usamos *perplexity*

Perplexity



Inversa de la probabilidad en un set de testeo, normalizada por la cantidad de palabras

$$PP(W) = \sqrt[N]{\prod_{i=1}^N \frac{1}{P(w_i|w_1 \dots w_{i-1})}}$$

$$\begin{aligned} \text{Perplexity}(C) &= \sqrt[N]{\frac{1}{\prod_{i=1}^m p(s_i)}} \\ &= 2^{\log_2 [\prod_{i=1}^m p(s_i)]^{-N}} \\ &= 2^{-\frac{1}{N} \log_2 [\prod_{i=1}^m p(s_i)]} \\ &= 2^{-\frac{1}{N} \sum_{i=1}^m \log_2 p(s_i)} \end{aligned}$$

La perplexity solo es comparable si dos modelos fueron entrenados con el mismo corpus

- Esto se debe al tratamiento que se le da a los <UNK>
- Si tenemos vocabulario chico y asignamos alta probabilidad a <UNK>, la PP va a ser muy alta

Detalles de implementación



- Usar log probability
- Marcar inicio y final de oración (<s> </s>)
- Cuidar el género de los textos
- Cuidar los dialectos de los textos
- Tratamiento de las palabras no vistas en el train
 - Marcar como <unk> en training todas las que tienen muy baja frecuencia
 - Tratar a todos los <unk> igual
- Tratamiento de probabilidades = 0 (smoothing)

Smoothing



Training set:

Cuando llegue a casa me tengo que acordar de darle de comer al gato
El perro me comió la tarea
Me olvidé de darle de comer al canario

Test set:

Cuando llegue a casa me tengo que acordar de darle de comer al perro

$$p(\text{perro} \mid \text{darle de comer al}) = ?$$

Smoothing: redistribución de las probabilidades

Smoothing



Laplace smoothing (add-one)

$$P_{\text{Laplace}}(w_i) = \frac{c_i + 1}{N + V}$$

Add-K smoothing

$$P_{\text{Add-k}}^*(w_n | w_{n-1}) = \frac{C(w_{n-1}w_n) + k}{C(w_{n-1}) + kV}$$

Backoff and Interpolation: usando menos contexto (n mas chico) cuando no existe el n-grama

¿Cómo generalizamos en N-gram?



Entrenamiento: "Cuando llegue a casa me tengo que acordar de darle de comer al gato"

¿Cuál es la probabilidad 5-gram de "perro" dada la siguiente frase?

Testeo: "Me estaba yendo a dormir y justo me acordé de darle la comida al ____"

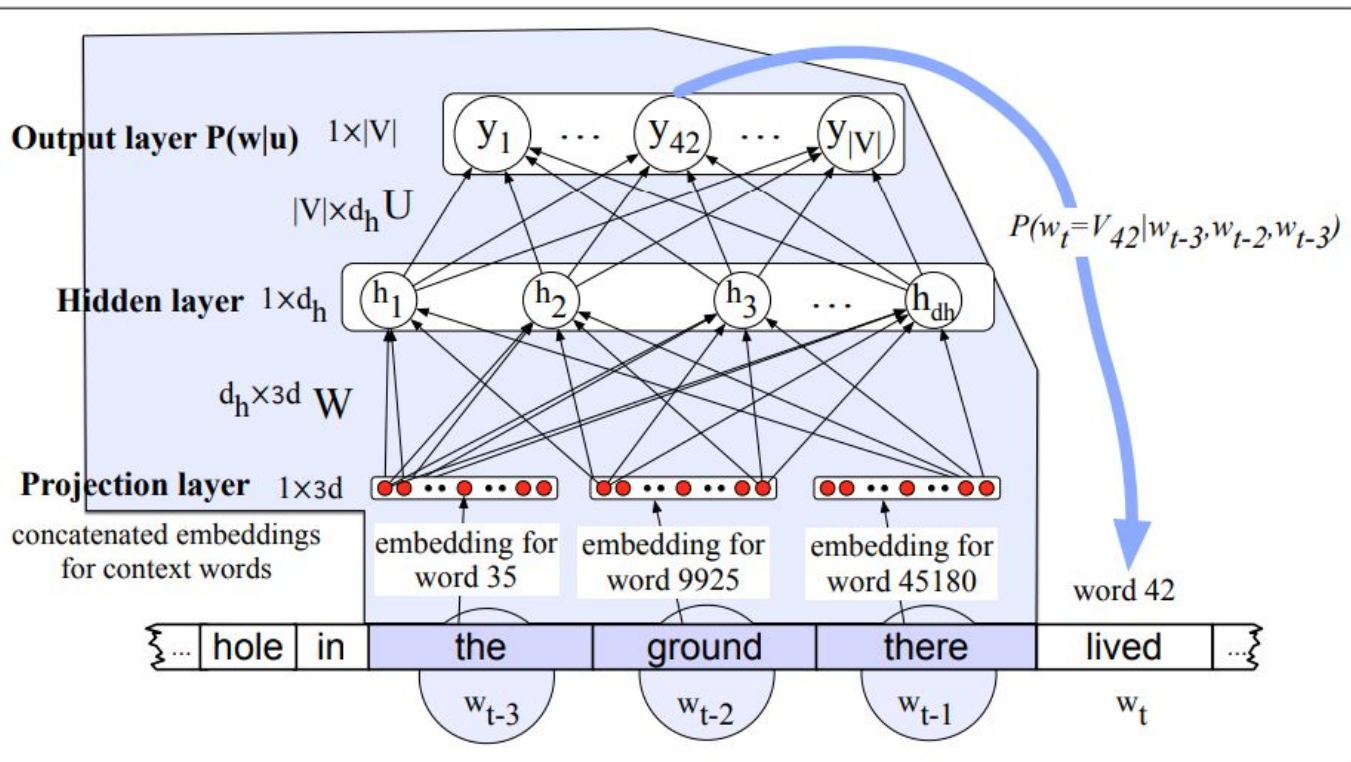
N-gram no puede generalizar a palabras similares.

Los embeddings y las redes neuronales (feed forward) nos pueden ayudar

Modelos de Lenguaje con redes

(Neural Language Models)

Neural LM

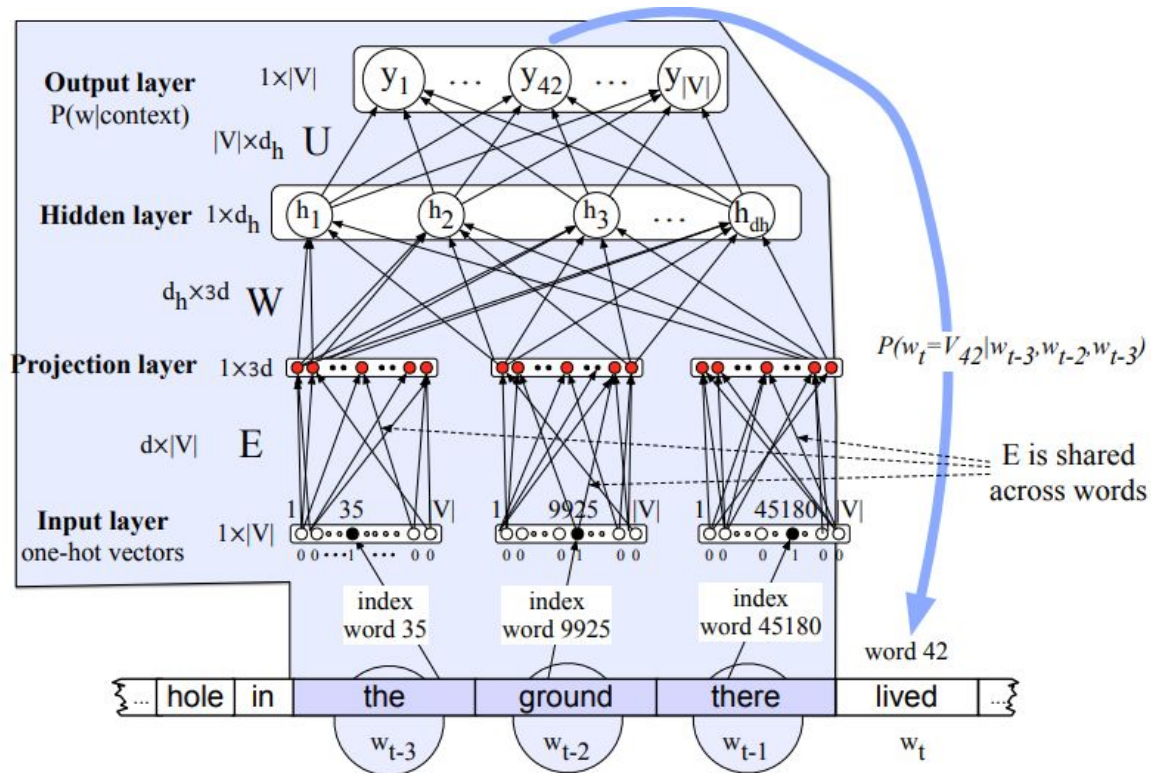


Podemos usar embeddings pre-entrenados

Estos embeddings ya tienen codificadas las relaciones semánticas

Pero no están calculados para optimizar la tarea para la cual creamos la red (ej. Predicción)

Neural LM



Para mejorar nuestra red podemos agregarle una capa donde se calcule el embedding.

La primera capa va a tomar los one-hot encodings de las palabras, pasa por la capa de embedding y luego a la capa oculta.

El backpropagation incluye a los embeddings, por lo que se calcular teniendo en cuenta la tarea

Ejemplos

Prediciendo la predictibilidad



¿Qué es la predictibilidad?

Predictibilidad: probabilidad de conocer una palabra antes de leerla, basado en su contexto precedente



¿Cómo medimos la predictibilidad?

Predictibilidad: probabilidad de conocer una palabra antes de leerla, basado en su contexto precedente

EI



¿Cómo medimos la predictibilidad?

Predictibilidad: probabilidad de conocer una palabra antes de leerla, basado en su contexto precedente

El relator



¿Cómo medimos la predictibilidad?

Predictibilidad: probabilidad de conocer una palabra antes de leerla, basado en su contexto precedente

El relator festejó



¿Cómo medimos la predictibilidad?

Predictibilidad: probabilidad de conocer una palabra antes de leerla, basado en su contexto precedente

El relator festejó el



¿Cómo medimos la predictibilidad?

Predictibilidad: probabilidad de conocer una palabra antes de leerla, basado en su contexto precedente

El relator festejó el gol



¿Cómo medimos la predictibilidad?

Predictibilidad: probabilidad de conocer una palabra antes de leerla, basado en su contexto precedente

El relator festejó el gol gritando



¿Cómo medimos la predictibilidad?

Predictibilidad: probabilidad de conocer una palabra antes de leerla, basado en su contexto precedente

El relator festejó el gol gritando “de



¿Cómo medimos la predictibilidad?

Predictibilidad: probabilidad de conocer una palabra antes de leerla, basado en su contexto precedente

El relator festejó el gol gritando “de qué



¿Cómo medimos la predictibilidad?

Predictibilidad: probabilidad de conocer una palabra antes de leerla, basado en su contexto precedente

El relator festejó el gol gritando “de qué planeta



¿Cómo medimos la predictibilidad?

Predictibilidad: probabilidad de conocer una palabra antes de leerla, basado en su contexto precedente

El relator festejó el gol gritando “de qué planeta viniste?”



¿Cómo medimos la predictibilidad?

Predictibilidad: probabilidad de conocer una palabra antes de leerla, basado en su contexto precedente

Problema:

- Medir la predictibilidad es muy costoso
- Es necesario medir la predictibilidad para cada corpus de texto por separado

Objetivos:

- Desarrollar un reemplazo computacional de una medida costosa
- Lograr una mejor comprensión de las pistas utilizadas para predecir



¿Cómo predecimos?



El relator festejó el gol gritando “de qué planeta viniste?”



¿Cómo predecimos?

Alta predictibilidad por relaciones gramaticales

El relator festejó el gol gritando “de qué planeta viniste?”

Alta predictibilidad por relaciones semánticas

Alta predictibilidad por relaciones mnemónicas



¿Podemos modelarla?

Objetivos

- Desarrollar un reemplazo computacional de una medida costosa
- Lograr una mejor comprensión de las pistas utilizadas para predecir

Metodología

- LSA [Bianchi, 2014 (JAIIO)] -> **Entrenado con P/12**
- Wrod2Vec [Bianchi, 2014 (Tesis)] -> **Entrenado con Wikipedia**
- N-gram [Kamienkowski, 2016 (JAIIO)] -> **Entrenado con cuentos**



¿Podemos modelarla?

Objetivos

- Desarrollar un reemplazo computacional de una medida costosa
- Lograr una mejor comprensión de las pistas utilizadas para predecir

Metodología

- LSA [Bianchi, 2014 (JAIIIO)] -> Entrenado con P/12
- Wrod2Vec [Bianchi, 2014 (Tesis)] -> Entrenado con Wikipedia
- N-gram [Kamienkowski, 2016 (JAIIIO)] -> Entrenado con cuentos

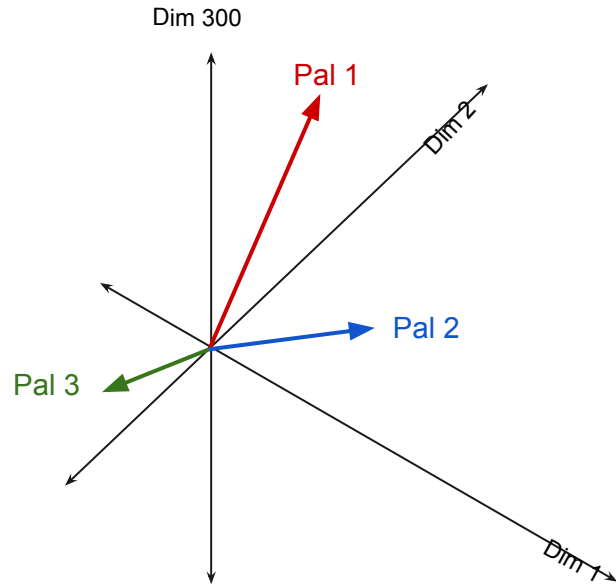
Unificar entrenamiento

2082 cuentos

107M pals



LSA y word2vec | Predicción



Dado un contexto y la palabra que lo continúa:

Cuando estaba como naturalista a bordo del Beagle, buque de la marina real, me impresionaron mucho ciertos hechos que se presentan en la distribución _____

geográfica

Calculo el vector promedio del contexto y la distancia coseno a la palabra target



n-grams | Predicción

La última pregunta se formuló
última pregunta se formuló por
pregunta se formuló por primera
se formuló por primera vez,
formuló por primera vez, medio
por primera vez, medio en
primera vez, medio en broma
vez, medio en broma, el
medio en broma, el 21
en broma, el 21 de mayo
broma, el 21 de mayo de
el 21 de mayo de
21 de mayo de 2061
de mayo de 2061, en
mayo de 2061, en momentos
de 2061, en momentos en
2061, en momentos en que
en momentos en que la
momentos en que la humanidad
en que la humanidad se



Dado un contexto y la palabra que lo continúa:

Cuando estaba como naturalista a bordo del Beagle, buque de la marina real, me impresionaron mucho ciertos hechos que se presentan en la distribución _____

geográfica

Calculo la probabilidad de co-ocurrencia de sus 4 palabras predecesoras.

Tiene que haber ocurrido esa secuencia en el entrenamiento. Si no, se usan suavizados

Bianchi et al. (2020) Scientific Reports



¿Qué es n-grams? | Predicción

Además de usar el 4-gram entrenado con el corpus de entrenamiento usamos un unigram caché:

- Funciona como una actualización sobre el texto leído
- Se mezcla con el 4-gram de forma lineal

$$P(\text{mezcla}) = (1-\lambda) * p(\text{nGram}) + \lambda * p(\text{caché})$$



¿Cómo sabemos si estamos estimando bien?

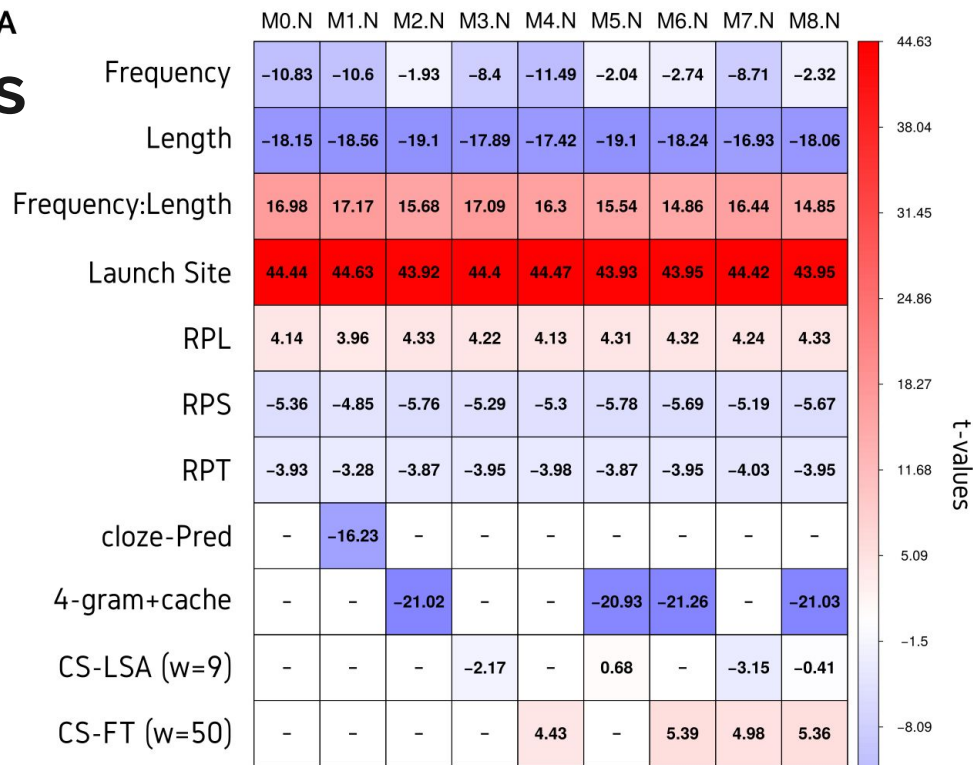
- Usamos Modelos Lineales Mixtos:
 - Variable Dependiente: movimientos oculares
 - Variables explicativas: long., freq., Pos. Linea, Pos.Texto, Pos.Oración, Repetición

Partimos de un modelo base (sin pred) -> Vamos incrementando la complejidad



Modelos

A



B

cloze-Pred (remef)	-	-	-9.47	-15.99	-16.55	-9.48	-9.87	-16.39	-9.87
--------------------	---	---	-------	--------	--------	-------	-------	--------	-------