



# Procesamiento de Lenguaje Natural

## Modelos de Lenguaje

Mauricio Toledo-Acosta  
[mauricio.toledo@unison.mx](mailto:mauricio.toledo@unison.mx)

Departamento de Matemáticas  
Universidad de Sonora



## Section 1

# Introducción



# Referencias

## Procesamiento de Lenguaje Natural

### Introducción

### Modelo de $n$ -gramas

### Generalizaciones

### Evaluando Modelos de Lenguaje

### Aspectos adicionales

### Algunas aplicaciones

**Chapter 3.** Jurafsky, D., Martin, J. H. (2019). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.



# Objetivo

Procesamiento  
de Lenguaje  
Natural

Introducción

Modelo de  
 $n$ -gramas

Generalizaciones

Evaluando  
Modelos de  
Lenguaje

Aspectos  
adicionales

Algunas  
aplicaciones

¿Cuál es la siguiente palabra?

Please turn your homework ...



# Objetivo

Procesamiento  
de Lenguaje  
Natural

Introducción

Modelo de  
 $n$ -gramas

Generalizaciones

Evaluando  
Modelos de  
Lenguaje

Aspectos  
adicionales

Algunas  
aplicaciones

¿Cuál es la siguiente palabra?

Please turn your homework ...

Los **modelos de lenguaje** son modelos que asignan una probabilidad a secuencias de palabras.



## Objetivo

Please turn your homework ...

Los **modelos de lenguaje** son modelos que asignan una probabilidad a secuencias de palabras.

- on guys all I of notice sidewalk three a sudden standing the
- all of a sudden I notice three guys standing on the sidewalk



5/32


$$P(\text{Eyes awe of an})$$

- Respuesta de preguntas
- Generación de texto





◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ↺ 🔍 ↻ 7/32


$$P(W) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2) \cdots P(w_n | w_1, w_2, \dots, w_{n-1})$$



# ¿Cómo se calculan las probabilidades?

its water is so transparent that the

Podemos estimar probabilidades en términos de conteos:

$$P(\text{the} \mid \text{its water is so transparent that}) = \frac{\text{contar}(\text{its water is so transparent that the})}{\text{contar}(\text{its water is so transparent that})}$$

Esto se llama *frecuencia relativa*.



*La distribución de probabilidad del valor futuro de una variable aleatoria depende únicamente de su valor presente, siendo independiente de la historia de dicha variable.*

Simplificamos cada término como

$$P(w_n | w_1 w_2 \dots w_{n-1}) \approx P(w_n)$$

$$P(w_n|w_1w_2...w_{n-1}) \approx P(w_n|w_{n-1})$$

$$P(w_n | w_1 w_2 \dots w_{n-1}) \approx P(w_n | w_{n-2} w_{n-1})$$

• • •



## Section 2

# Modelo de $n$ -gramas



# Modelo de $n$ -gramas

## Procesamiento de Lenguaje Natural

### Introducción

### Modelo de $n$ -gramas

### Generalizaciones

### Evaluando Modelos de Lenguaje

### Aspectos adicionales

### Algunas aplicaciones

Un  **$n$ -grama** es una secuencia de  $N$  palabras: un 2-grama (o bigrama) es una secuencia de dos palabras como *please turn*, *turn your your homework*, y un 3-grama (o trigram) es una secuencia de tres palabras como *please turn your* o *turn your homework*.

El **modelo de  $n$ -gramas** estima la probabilidad de una palabra dada una secuencia de palabras y estima la probabilidad de una secuencia de palabras.

Al decir  $n$ -gramas nos referimos a las secuencias o al modelo predictivo que asigna probabilidades.



- N-grams help capture the contextual information and semantics within a sequence of words, providing a more nuanced understanding of language.
- In information retrieval tasks, N-grams assist in matching and ranking documents based on the relevance of N-gram patterns.
- N-grams serve as powerful features in text classification and sentiment analysis, capturing meaningful patterns that contribute to the characterization of different classes or sentiments.









the in	rain Spain	in falls
Spain mainly	falls on	mainly the
on plain		



17/32





- La **evaluación extrínseca** es la evaluación del desempeño del modelo en la tarea particular para la cual está siendo entrenado. La evaluación extrínseca es la única forma de saber si una mejora concreta de un componente va a ayudar realmente a la tarea que se está realizando.
- La **evaluación intrínseca** mide la calidad del modelo independientemente de la tarea o aplicación del modelo. Algunos ejemplos son: Entropía, Perplejidad, etc.


$$\begin{aligned} \text{Pp}(W) &= \sqrt[N]{\frac{1}{P(w_1 w_2 \cdots w_N)}} \\ &= \sqrt[N]{\prod_i \frac{1}{P(w_i | w_{i-1})}} \end{aligned}$$

$W$  es la secuencia entera de palabras de un conjunto de prueba.



Considerar textos escritos en AAVE (African American Vernacular English):



## Perplejidad

## Procesamiento de Lenguaje Natural

## Generalizaciones

- Al calcular la perplejidad, el modelo de  $n$ -gramas debe construirse sin ningún conocimiento del conjunto de prueba. De otra forma la perplejidad puede ser artificialmente baja.
- La perplejidad de dos modelos lingüísticos sólo es comparable si utilizan vocabularios idénticos.
- Una mejora en la perplejidad (m. intrínseca) no garantiza una mejora del rendimiento de una tarea de PLN como el reconocimiento del habla o la traducción automática (m. extrínseca). Sin embargo, la perplejidad suele estar correlacionada con dichas mejoras.







# Suavizado: El papel del corpus de entrenamiento

## Procesamiento de Lenguaje Natural

### Introducción

### Modelo de $n$ -gramas

### Generalizaciones

### Evaluando Modelos de Lenguaje

### Aspectos adicionales

### Algunas aplicaciones

1 gram	-To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have -Hill he late speaks; or! a more to leg less first you enter
2 gram	-Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow. -What means, sir. I confess she? then all sorts, he is trim, captain.
3 gram	-Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, 'tis done. -This shall forbid it should be branded, if renown made it empty.
4 gram	-King Henry. What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv'd in; -It cannot be but so.

1 gram	Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives
2 gram	Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her
3 gram	They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions



denied the offer: 0  
denied the loan: 0



# Suavizado: Palabras desconocidas

## Procesamiento de Lenguaje Natural

### Introducción

### Modelo de *n*-gramas

### Generalizaciones

### Evaluando Modelos de Lenguaje

### Aspectos adicionales

### Algunas aplicaciones

¿Qué pasa con palabras que nunca ha visto en el entrenamiento?

En un sistema de **vocabulario cerrado** el conjunto de prueba no contiene palabras desconocidas. En un sistema de **vocabulario abierto**, tenemos que lidiar con palabras que no hemos visto antes, a las que llamaremos **palabras fuera de vocabulario (OOV)**. Podemos lidiar con estas palabras desconocidas añadiendo una pseudopalabra llamada **<UNK>**.



- Convertir un sistema abierto en uno cerrado:
  - Escoger un vocabulario fijo.
  - Convertir cualquier palabra OOV en  $\langle \text{UNK} \rangle$ .
  - Estimar las probabilidades para  $\langle \text{UNK} \rangle$  de la forma usual, como si fuera una palabra *normal*.
- Crear un vocabulario fijo implícito. Reemplazamos palabras por  $\langle \text{UNK} \rangle$  en el entrenamiento basándonos en su frecuencia.



- Add- $k$  smoothing
- Kneser-NeySmoothing
- ...

	i	want	to	eat	chinese	food	lunch	spend
i	6	828	1	10	1	1	1	3
want	3	1	609	2	7	7	6	2
to	3	1	5	687	3	1	7	212
eat	1	1	3	1	17	3	43	1
chinese	2	1	1	1	1	83	2	1
food	16	1	16	1	2	5	1	1
lunch	3	1	1	1	1	2	1	1
spend	2	1	2	1	1	1	1	1



## Section 6

# Algunas aplicaciones









## 32/32