



Procesamiento de Lenguaje Natural

Modelos de Lenguaje

Mauricio Toledo-Acosta
mauricio.toledo@unison.mx

Departamento de Matemáticas
Universidad de Sonora



Section 1

Introducción



Referencias

- **Chapter 3.** Jurafsky, D., Martin, J. H. (2019). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.
- **Chapter II.6.** Eisenstein, J. (2018). Natural language processing. Jacob Eisenstein.



Objetivo

Procesamiento
de Lenguaje
Natural

Introducción

Modelo de
 n -gramas

Generalizaciones

Evaluando
Modelos de
Lenguaje

Aspectos
adicionales

Algunas
aplicaciones

¿Cuál es la siguiente palabra?

En el parque, los niños juegan con ...



Objetivo

Procesamiento
de Lenguaje
Natural

Introducción

Modelo de
 n -gramas

Generalizaciones

Evaluando
Modelos de
Lenguaje

Aspectos
adicionales

Algunas
aplicaciones

¿Cuál es la siguiente palabra?

En el parque, los niños juegan con ...

¿Cuál es la siguiente palabra?

El agua hierve a 100 grados ...



- *café el en mesa la sobre libro un dejé olvidé y*



- *café el en mesa la sobre libro un dejé olvidé y*
- *dejé un libro sobre la mesa en el café y lo olvidé*



6/41


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

- 7/41



◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ↺ 🔍 ↻ 8/41


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



◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ↺ 🔍 ↻ 10/41



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.

Podemos hacer las siguientes aproximaciones:

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

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

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

• • •



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



Usos del modelo de n -gramas

Procesamiento de Lenguaje Natural

Introducción

Modelo de n -gramas

Generalizaciones

Evaluando Modelos de Lenguaje

Aspectos adicionales

Algunas aplicaciones

Hay dos principales utilidades de este modelo:

- 1 **Tareas secuenciales:** Modelar la dependencia entre elementos consecutivos en una secuencia (palabras, caracteres, etc.). Por ejemplo: Generación de texto, corrección ortográfica, etc.
- 2 **Tareas no secuenciales:** Representar el contenido sin considerar el orden estricto, enfocándose en la presencia o frecuencia de patrones. Ejemplos: Tareas de clasificación, recuperación de información, etc.



15/41



Cálculo de probabilidades condicionales para 2-gramas

El cálculo de probabilidades en modelos basados en 2-gramas se realiza mediante la siguiente fórmula:

$$P(w_i | w_{i-1}) = \frac{\text{Frecuencia}(w_{i-1}, w_i)}{\sum_j \text{Frecuencia}(w_{i-1}, w_j)}$$

- $\text{Frecuencia}(w_{i-1}, w_i)$: Número de veces que la palabra w_i aparece inmediatamente después de w_{i-1} en el corpus.
- $\sum_j \text{Frecuencia}(w_{i-1}, w_j)$: Suma de todas las frecuencias de palabras que siguen a w_{i-1} .

Ejemplo: Si en el corpus tenemos:

- $\text{Frecuencia}(\text{"me"}, \text{"gusta"}) = 50$,
- $\text{Frecuencia total para "me"} = 61$,

$$P(\text{gusta} | \text{me}) = \frac{50}{61} \approx 0.82$$



Ejemplo: Probabilidades

Procesamiento de Lenguaje Natural

Introducción

Modelo de *n*-gramas

Generalizaciones

Evaluando Modelos de Lenguaje

Aspectos adicionales

Algunas aplicaciones

	me	gusta	comer	tacos	tamales	mucho	hoy
me	0.082	0.820	0.000	0.033	0.016	0.000	0.049
gusta	0.033	0.000	0.667	0.017	0.067	0.100	0.000
comer	0.000	0.032	0.000	0.484	0.403	0.000	0.000
tacos	0.020	0.000	0.060	0.000	0.100	0.020	0.020
tamales	0.000	0.029	0.086	0.200	0.000	0.029	0.000
mucho	0.211	0.316	0.000	0.000	0.053	0.000	0.421
hoy	0.167	0.000	0.083	0.083	0.000	0.667	0.000

Observa las sumatorias por fila.



- Los n-gramas solo capturan dependencias locales.
- Requiere grandes corpus para evitar probabilidades cero en combinaciones raras.
- No entiende significado, solo patrones estadísticos.



- Los n -gramas ayudan a capturar la información contextual y la semántica dentro de una secuencia de palabras, proporcionando una comprensión más matizada del lenguaje.
- En tareas de recuperación de información (information retrieval), los n -gramas ayudan a emparejar y clasificar documentos según la relevancia de los patrones de n -gramas.
- Los n -gramas sirven como características poderosas en la clasificación de texto y el análisis de sentimientos, capturando patrones significativos que contribuyen a la caracterización de diferentes clases o sentimientos.



Algunas desventajas

- En general es un modelo de lenguaje insuficiente. Por ejemplo, no captura dependencias lejanas

The computer which I had just put into the machine room on the fifth floor crashed.

En el caso anterior el bigrama *computer-crashed* puede ser un bigrama importante.

- El lenguaje es creativo, todo el tiempo se crean asociaciones nuevas, y no siempre podremos contar frases enteras.
- Problemas de escalabilidad y almacenamiento. A medida que aumenta el valor de n , el número de combinaciones posibles crece exponencialmente.



Actividad en clase

Procesamiento de Lenguaje Natural

Introducción

Modelo de n -gramas

Generalizaciones

Evaluando Modelos de Lenguaje

Aspectos adicionales

Algunas aplicaciones

En equipos de, máximo 3 personas, haz la tabla de frecuencias y probabilidades para el siguiente texto:

El aprendizaje automático está transformando la industria. El aprendizaje está transformando los modelos de lenguaje. Los modelos de lenguaje son un claro de revolución tecnológica. La industria de la revolución tecnológica está cambiando con el aprendizaje automático. El aprendizaje está transformando todo.



Section 3

Generalizaciones



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



n -gramas sintácticos

Procesamiento
de Lenguaje
Natural

Introducción

Modelo de
 n -gramas

Generalizaciones

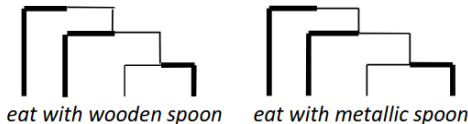
Evaluando
Modelos de
Lenguaje

Aspectos
adicionales

Algunas
aplicaciones

A diferencia de los n -gramas donde las subsecuencias se toman en el orden en el que aparecen en el texto, en los **n -gramas sintácticos** los vecinos se toman siguiendo las relaciones sintácticas de los árboles de dependencia sintáctica.

eat with wooden spoon eat with metallic spoon



¿Cuántos bigramas y bigramas sintácticos en común tienen?



Section 4

Evaluando Modelos de Lenguaje



Evaluando Modelos de Lenguaje

Procesamiento de Lenguaje Natural

Introducción

Modelo de n -gramas

Generalizaciones

Evaluando Modelos de Lenguaje

Aspectos adicionales

Algunas aplicaciones

Hay dos maneras de evaluar un modelo de Lenguaje:

- 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.



Perplejidad

Procesamiento de Lenguaje Natural

Introducción

Modelo de n-gramas

Generalizaciones

Evaluando Modelos de Lenguaje

Aspectos adicionales

Algunas aplicaciones

La **perplejidad** es una métrica para evaluar el rendimiento de un modelo de lenguaje. Mide la incertidumbre del modelo para predecir la próxima palabra en una secuencia. Cuanto menor sea la perplejidad, mayor será la capacidad del modelo para predecir la palabra siguiente.

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

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



La perplejidad se puede interpretar como **el número promedio de opciones que el modelo está considerando** para la siguiente palabra.

- **Perplejidad baja:** El modelo está más seguro, tiene menos incertidumbre
- **Perplejidad alta:** El modelo está más confundido, considera muchas opciones

Ejemplo extremo:

- Si el modelo siempre asigna probabilidad 1.0 a la palabra correcta: $P_p = 1$ (perfecto)
- Si el modelo considera 100 palabras con igual probabilidad: $P_p = 100$ (muy confundido)

Buscamos perplejidad baja



Perplejidad: Ejemplo Numérico

Procesamiento
de Lenguaje
Natural

Introducción

Modelo de
 n -gramas

Generalizaciones

Evaluando
Modelos de
Lenguaje

Aspectos
adicionales

Algunas
aplicaciones

Calculemos la perplejidad de un bigrama simple en la oración de prueba: <s> el gato come <e>

Supongamos que nuestro modelo bigrama asigna estas probabilidades:

$$P(\text{el} | < s >) = 0.5$$

$$P(\text{gato} | \text{el}) = 0.25$$

$$P(\text{come} | \text{gato}) = 0.2$$

$$P(< e > | \text{come}) = 0.8$$

Calculamos:

$$P_p = \sqrt[4]{\frac{1}{0.5 \times 0.25 \times 0.2 \times 0.8}} = \sqrt[4]{\frac{1}{0.02}} = \sqrt[4]{50} \approx 2.66$$

El modelo considera en promedio ~ 2.66 opciones por palabra.





- 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 clasificación (m. extrínseca). Sin embargo, la perplejidad suele estar correlacionada con dichas mejoras.



Section 5

Aspectos adicionales



Si un n -grama aparece un número suficiente de veces, podemos tener una buena estimación de su probabilidad. Sin embargo, es probable que falten algunas secuencias de palabras perfectamente aceptables. Tendremos casos de n -gramas de probabilidad 0 que deberían tener probabilidad distinta de 0.

denied the allegations:	5
denied the speculation:	2
denied the rumors:	1
denied the report:	1

denied the offer: 0
denied the loan: 0



Suavizado

Procesamiento de Lenguaje Natural

Introducción

Modelo de n -gramas

Generalizaciones

Evaluando Modelos de Lenguaje

Aspectos adicionales

Algunas aplicaciones

En los modelos de lenguaje estadísticos, como los n -gramas, a menudo nos encontramos con una secuencia de palabras que nunca apareció en el texto de entrenamiento. Si le asignamos una probabilidad de 0, el modelo se *rompe* porque cualquier oración que contenga esa secuencia tendrá una probabilidad total de 0. El suavizado es la técnica para evitar este problema, asignando una pequeña parte de la probabilidad a eventos no vistos.



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 $\langle \text{UNK} \rangle$.



Palabras desconocidas

Procesamiento de Lenguaje Natural

Introducción

Modelo de n -gramas

Generalizaciones

Evaluando Modelos de Lenguaje

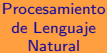
Aspectos adicionales

Algunas aplicaciones

Hay dos estrategias:

- 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.





Section 6

Algunas aplicaciones



Google n -grams

Procesamiento de Lenguaje Natural

Introducción

Modelo de n -gramas

Generalizaciones

Evaluando Modelos de Lenguaje

Aspectos adicionales

Algunas aplicaciones

- El **Visor de n -gramas de Google** es un motor de búsqueda que traza las frecuencias de n -gramas encontrados en fuentes impresas publicadas entre 1500 y 2022.
Algunos ejemplos: **1**, **2**, **3**
- **Google Research** puso disponible un **corpus grande** de n -gramas. Incluye n -gramas que ocurren al menos 40 veces en una secuencia de 1,024,908,267,229 palabras.



◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ↺ 🔍 ↻ 40/41



◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ ↺ 🔍 ↻ 41/41