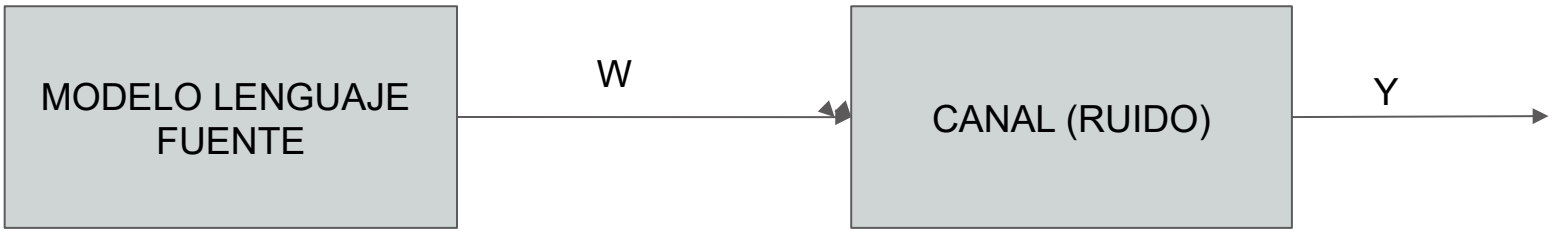
Introducción al Procesamiento de Lenguaje Natural

Grupo PLN - INCO

Modelos de Lenguaje

APLICACIÓN	SEÑAL	OBJETIVO
RECONOCIMIENTO DE HABLA	SEÑAL ACÚSTICA Y	VERSION STRING W DE Y
TRADUCCIÓN AUTOMÁTICA ESPAÑOL-INGLÉS	STRING EN ESPAÑOL	STRING EN INGLÉS
CORRECCIÓN ORTOGRÁFICA	TEXTO CON ERRORES	TEXTO SIN ERRORES
OPTICAL CHARACTER RECOGNITION (OCR)	IMAGEN DE UN TEXTO (PIXELS)	TEXTO CON CARACTERES INTERPRETABLES
GENERACIÓN DE TEXTO (P.EJ., RESPUESTAS A PREGUNTAS)	REPRESENTACIÓN SEMÁNTICA INTERNA	TEXTO GENERADO

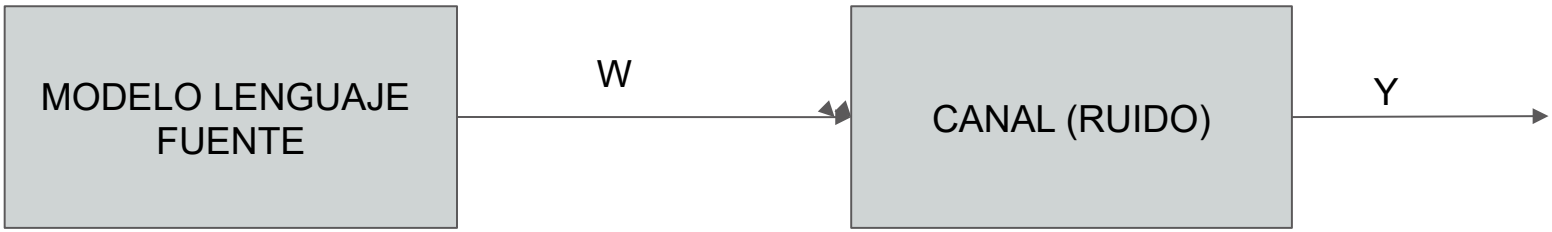
Modelos de Lenguaje



P(W) X P(Y | W) = P(W,Y)

El objetivo es determinar W a partir de Y !!

Modelos de Lenguaje

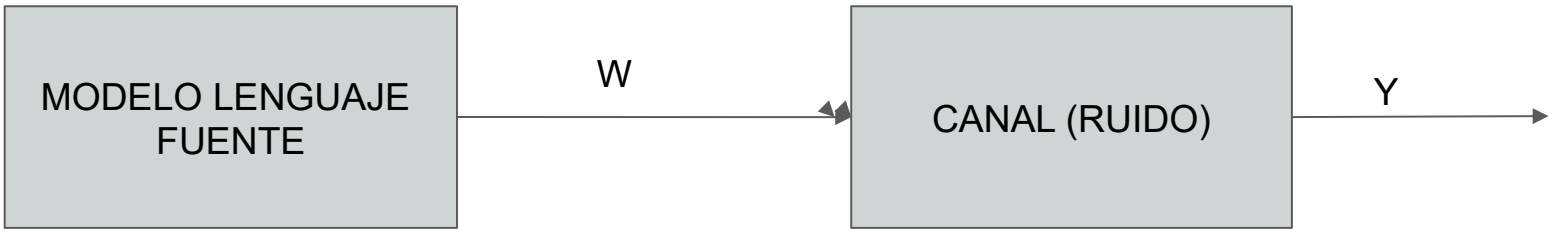


P(W) X P(Y | W) = P(W,Y)

El objetivo es determinar W a partir de Y !!

En corrección de errores, suponemos que hay un texto fuente correcto W que se distorsiona por un canal ruidoso.

Modelos de Lenguaje



P(W) X P(Y | W) = P(W,Y)

El objetivo es determinar W a partir de Y !!

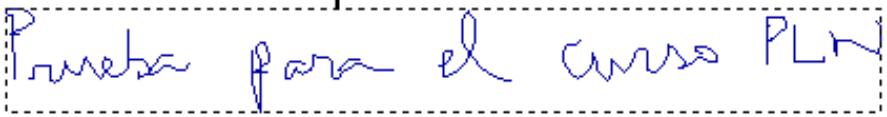
En corrección de errores, suponemos que hay un texto fuente correcto W que se distorsiona por un canal ruidoso.

En traducción español - inglés, hay un texto original en inglés del cual tenemos una versión defectuosa que es el texto en español.

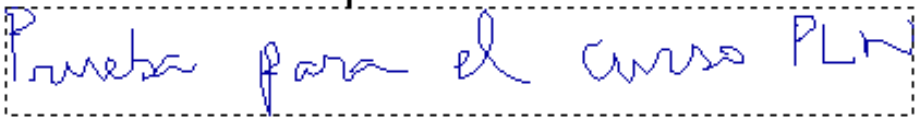
Modelos de Lenguaje

Reconocimiento de escritura.

Prneba para el cnrso PLN



Prueba para el curso PLN



(Ejemplos gentileza de Gustavo Crispino – Vision Objects – Nantes – France)

Modelos de Lenguaje

Reconocimiento de escritura

$P(W \mid Y) \propto P(Y \mid W)$ **P(W)** **(*)**

Modelos de Lenguaje

Reconocimiento de escritura

$P(W|Y) \propto P(Y|W)$ **P(W)** **(*)**

(*) Bayes!

Modelos de Lenguaje

- Se necesita el modelo del canal y el de la fuente.

Modelos de Lenguaje

- Se necesita el modelo del canal y el de la fuente.
- Nos ocuparemos del modelo de la FUENTE.

Modelos de Lenguaje

- Se necesita el modelo del canal y el de la fuente.
- Nos ocuparemos del modelo de la FUENTE.
- Veremos modelos probabilísticos *(No es la única opción, podría ser un modelo “duro”, generar el lenguaje por una CFG)*.

Modelos de Lenguaje

- Se necesita el modelo del canal y el de la fuente.
- Nos ocuparemos del modelo de la FUENTE.
- Veremos modelos probabilísticos *(No es la única opción, podría ser un modelo "duro", generar el lenguaje por una CFG)*.
- Es imposible recuperar W exitosamente en todos los casos, trabajar con probabilidades permite minimizar errores.

Modelo Probabilista de Lenguaje

Objetivo : Probabilidad de una oración o secuencia de palabras W.

$$P(W) = P(w_1, w_2, w_3, w_4, w_5 \ldots w_n)$$

Modelo Probabilista de Lenguaje

Tarea relacionada : probabilidad de la palabra siguiente.

$$P(w_5 | w_1, w_2, w_3, w_4)$$

Modelo Probabilista de Lenguaje

¿ Cómo computar P(W) ?

- Imposible basarse en frecuencias relativas para oraciones enteras.
- ¿Por qué ?

Regla de la cadena

Definición de probabilidad condicional :

$$P(A \mid B) = P(A, B) / P(B)$$

Equivale a $$P(A, B) = P(A \mid B) P(B)$$

Regla de la cadena

En el caso general

$$P(x_1,x_2,x_3,\ldots,x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)\ldots P(x_n|x_1,\ldots,x_{n-1})$$

$$P(w_1 w_2 \ldots w_n) = \prod_i P(w_i \mid w_1 w_2 \ldots w_{i-1})$$

Predicción de palabras

Predecir la siguiente palabra a partir de las anteriores es todo lo que precisamos ...

El Inumet emitió un pronóstico ...

A raíz de estos fenómenos se sucederán tormentas ...

rachas de viento fuerte de componente ...

Predicción de palabras

El Inumet emitió un pronóstico **especial**

A raíz de estos fenómenos se sucederán tormentas **fuertes**

rachas de viento fuerte de componente **sudoeste**

Probabilidad de una oración

P(viento fuerte de componente sudoeste) = ?

P(viento fuerte de componente sudoeste)= P(viento)P(fuerte|viento)P(de|viento fuerte)P(componente| viento fuerte de)P(sudoeste|viento fuerte de componente)

Regla de la Cadena

No se puede usar tal cual !!

Estimación de probabilidades

La probabilidad se estima a partir de las frecuencias de ocurrencias en un gran corpus (Principio de Máxima Verosimilitud: ajustar lo mejor posible a los datos).

Estimación de la probabilidad

¿Cómo podemos computar la probabilidad $P(w_n|w_1^{n-1})$?

No es manejable si no restringimos la historia

$W_1, W_2, \ldots, W_{n-N}, w_{n-N+1}, \ldots w_{n-1}, w_n$

Hipótesis markoviana

Bigrama: $P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-1}^{n-1})$

Trigrama: $P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-2}^{n-1})$

N-Grama: $P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-N+1}^{n-1})$

Bigrama: $P(w_1^n) = \prod_{k=1}^{n} P(w_k|w_{k-1})$

Ngramas

Un modelo de ngramas (modelo de lenguaje) intenta predecir la próxima palabra de una oración a partir de las N-1 anteriores.

El orden importa:

- de un vuelo de flamencos quemando un horizonte de bañados (Cortázar/La vuelta al día en ochenta mundos)
- vuelo de quemando flamencos un de un de bañados horizonte
- de un vuelo de bañados quemando un horizonte de flamencos

Ngramas

¿Qué elementos vamos a contar para modelar el lenguaje?

Con alivio, con humillación, con terror, comprendió que él también era una apariencia, que otro estaba soñándolo.

¿17 o 22 palabras?

¿Y en un corpus oral?

Eh.. en reali- en realidad yo creía.. creía que era más sencillo.

Disfluencias: fragmentos, rellenos, repetición de palabras…

Mayúsculas, formas flexionadas…

Ngramas

Caso más simple: unigramas

$$P(w_1 w_2 \ldots w_n) \approx \prod_i P(w_i)$$

Hipótesis de independencia demasiado fuerte !!!

Ngramas

¿Cuál de las siguientes tendrá mayor probabilidad con el modelo de unigramas: ?

Me gustan los helados.
El tren no pasó.
El el el el.

Ngramas, problema

Utilizar más que trigramas es muy complejo.

Pero hay dependencias de larga distancia que no se pueden representar

El libro que María compró la semana pasada en la exposición es interesante.

Los libros que María compró la semana pasada en la exposición son interesantes.

Cantidad de palabras

¿Cuántas palabras hay en los idiomas?

- Oxford English Dictionary: 290.500
- Trésor de la langue française: 54.280
- Diccionario de la RAE (22ª ed.): 88.431

¿Y en un CORPUS?

- Brown Corpus -> 1.000.000
- CREA (RAE) -> 250.000.000 (en revisión)

Distinguir entre el número de Tokens (T) y el número de palabras distintas (V).

Estimación de probabilidades

¿Cómo estimamos los bigramas?

Se utiliza un Estimador de Máxima Verosimilitud (o método de frecuencias relativas)

$$P(w_n|w_{n-1}) = \frac{c(w_{n-1}w_n)}{\sum_w c(w_{n-1}w))}$$

Estimación de probabilidades

¿Cómo tomar en cuenta la primera y última palabra de una oración? Agregamos símbolos <s> y </s>.

Simplificación de la fórmula:

$$P(w_n|w_{n-1}) = \frac{c(w_{n-1}w_n)}{c(w_{n-1})}$$

Bigramas

<s> juan abrió la puerta </s>

<s> el viento abrió la puerta </s>

<s> enero abrió limones en tus mejillas nuevas </s>

<s> juan recoge limones </s>

1. P(<s> juan abrió limones </s>)
2. P(<s> enero abrió la puerta </s>)
3. P(<s> juan come </s>)
4. P(<s> en la puerta </s>)

Bigramas

P(<s> juan abrió limones </s>) =

P(juan | <s>) . P(abrió | juan) . P(limones | abrió) . P(</s> | limones)

= [c(<s> juan) / c(<s>)] . [c(juan abrió) / c(juan)] .
[c(abrió limones) / c(abrió)] . [c(limones </s>) / c(limones)]
= 2/4 . 1/2 . 1/3 . 1/2 = 1/24 ≈ 0.042

P(<s> enero abrió la puerta </s>) = 1/6 ≈ 0.17

P(<s> juan come </s>) = 0

P(<s> en la puerta </s>) = 0

Ngramas: determinar N

A mayor valor del N, mejores resultados. Sin embargo con N=3 (trigramas) generalmente da muy buenos resultados.

Trigramas: se agregan dos símbolos para inicio de oración y dos para fin de oración.

P.ej: P(juan | <s> <s>)

Corpus

Las probabilidades de un modelo de Ngramas se obtienen a partir de un corpus de entrenamiento.

Corpus

Idea: utilizar dos corpus, uno de entrenamiento y otro de prueba. Dado un problema:

- se recopila un conjunto de textos relevantes
- se divide en un corpus de entrenamiento (CE) y en un corpus de prueba (CP)
- se entrena el modelo
- se lo testea para evaluar el modelo

¿Cómo dividimos?

- queremos entrenar lo máximo posible
- un corpus de prueba pequeño puede ser no significativo
- en general: 90% para el CE y 10% para el CP

Evaluación de modelos

La manera “correcta” de evaluar un modelo es empíricamente “in vivo”.

P.ej: Reconocimiento de habla: tomamos dos modelos, los incluimos en una aplicación y medimos el que ayuda a reconocer mejor.

Problema: puede ser costoso: horas o días de procesamiento de un corpus oral.

Evaluación de modelos

Perplejidad (PP):

de un modelo de lenguaje,

en un conjunto de testeo $CT=W_1W_2W_3 \ldots W_{n,}$

es la inversa de la probabilidad del conjunto de testeo, normalizada por la cantidad de elementos del conjunto

$$PP(CT) = P(w_1w_2w_3 \ldots w_n)^{-1/n}$$

Evaluación de modelos

Perplejidad (PP):

$$PP(CT) = P(w_1 w_2 w_3 \ldots w_n)^{-1/n}$$

Nos interesa maximizar la probabilidad de CT, equivale a minimizar la perplejidad.

Evaluación de modelos

Perplejidad (PP):

$$PP(CT) = P(w_1 w_2 w_3 \ldots w_n)^{-1/n}$$

Se usa en general la secuencia entera de palabras del test set, incluyendo separadores entre oraciones.

Evaluación de modelos

Perplejidad (PP):

$$PP(CT) = P(w_1 w_2 w_3 \ldots w_n)^{-1/n}$$

Paea computar esta probabilidad conjunto volvemos a usar la regla de la cadena !! (bigramas)

Evaluación de modelos

Ejemplo del libro:

- Entrenamiento de unigramas, bigramas y trigramas en un corpus de artículos periodísticos del Wall Street Journal (WSJ) de 38 millones de palabras.
- Utilización de un vocabulario de 19979 palabras.
- Se probaron los modelos sobre un corpus de prueba de 1.5 millones de palabras.
- Se computó la perplejidad para cada modelo.

N-gram Order	Unigram	Bigram	Trigram
Perplexity	962	170	109

Problemas

¡Probabilidades nulas!

Palabras no existentes:

- Se crea un vocabulario fijo.
- En el corpus se sustituyen las palabras desconocidas por una especial (p.ej: <UNK>).

Ngramas que no ocurren: técnicas de suavizado.

Suavizado de Ngramas

Idea simple: agregar 1 a todos los contadores.

Caso de unigramas

Asumiendo un corpus con T tokens y V palabras distintas:

$$P(w_i) = \frac{c(w_i))}{T} \longrightarrow P(w_i) = \frac{c(w_i) + 1}{T + V}$$

Suavizado de Ngramas

Ejemplo: “Esta es la historia de un hombre y la ciudad que él creó”

c(esta)=1 c(la)=2 c(quiso)=0

p(esta)=1/13(0.0833), p(la)=2/13 (0.1538), p(quiso)=0

Laplace:

p’(esta)=2/25 (0.08), p’(la)=3/25 (0.12), p’(quiso)=1/25 (0.04)

c*(esta)=2*(12/25)=0.96, c*(la)=1.44, c*(quiso)=0.48

d(esta)=0.96/1=0.96, d(la)=1.44/2=0.72, d(quiso)=...!

Suavizado de Ngramas

Esta técnica no funciona muy bien y no suele usarse en la práctica.

"Mueve" demasiada masa de probabilidad hacia los Ngramas con probabilidad cero.

Se puede utilizar una fracción δ en lugar de 1. Hay que calcularla (¿Cómo buscamos el δ óptimo?)

Interpolación y Backoff

Otra manera de resolver el problema de las probabilidades nulas.

Objetivo: computar la probabilidad

$$P(w_n \mid w_{n-1}\ w_{n-2})$$

Problema: no existen ocurrencias de $w_{n-2}\ w_{n-1}\ w$ en el texto.

Heurística: estimar $P(w_n \mid w_{n-1}\ w_{n-2})$ a partir de $P(w_n \mid w_{n-1})$

Interpolación y Backoff

Backoff: sólo se retrocede a un orden menor de N-gramas cuando no se tiene evidencia de un cierto orden mayor. P.ej: Para un trigrama que no ocurre en el texto se toma la probabilidad del bigrama sufijo.

Interpolación: se combinan las probabilidades de diferentes Ngramas para obtener la nueva probabilidad.

Interpolación y Backoff

Ejemplo: interpolación lineal simple

$$
\begin{aligned}
P'(w_n|w_{n-1}w_{n-2}) = \lambda_1 P(w_n|w_{n-1}w_{n-2}) \\
+ \lambda_2 P(w_n|w_{n-1}) \\
+ \lambda_3 P(w_n)
\end{aligned}
$$

Interpolación y Backoff

Ejemplo: interpolación condicionada por el contexto.

$$P'(w_n|w_{n-1}w_{n-2}) = \lambda_1(w_{n-2}^{n-1})P(w_n|w_{n-1}w_{n-2}) + \lambda_2(w_{n-2}^{n-1})P(w_n|w_{n-1}) + \lambda_3(w_{n-2}^{n-1})P(w_n)$$

Idea clave: si, p.ej., tenemos un conteo fuerte de un bigrama, podemos ponderar con más fuerza los trigramas basados en ese bigrama.

Los λ_i se estiman a partir de un corpus (*held-out corpus*, aprox. 10% del total) y forman parte del proceso de tuning del modelo.

Referencias

J.Martin & D.Jurafsky. Speech and Language Processing. Capítulo 4 (3era edición)