

Procesamiento de Lenguaje Natural

Apunte de Clases (Borrador)

Felipe Bravo Márquez

Felipe Bravo Márquez

Ilustración Portada por Paulette Filla

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN, UNIVERSIDAD DE CHILE

GITHUB.COM/DCCUCHILE/CC6205

Apuntes de clases del curso de Procesamiento de Lenguaje Natural de la Universidad de Chile.

El formato del apunte fue tomado del template de Jasmine Hao.

Borrador, 2 de noviembre de 2023



Índice general

0.1	Interpolación Lineal	3
0.1.1	Estimación de los Valores λ	4
0.2	Modelos de Descuento (Katz Back-Off)	4
0.3	Historia	7
0.4	Conclusiones	8

0.1 Interpolación Lineal

La principal idea detrás de la técnica de interpolación lineal es combinar las distribuciones de probabilidad estimadas por modelos de n-gramas con las obtenidas a través de modelos de órdenes inferiores, como bigramas y unigramas. Esta estrategia ofrece la ventaja de incorporar información de n-gramas de órdenes más bajos, permitiendo abordar la indefinición de probabilidades en palabras no observadas durante el entrenamiento. Pues basta que una oración tenga un sólo trígrama no observado en entrenamiento para recibir una probabilidad de cero por el efecto multiplicativo de la regla de la cadena.

En el modelo interpolado, se ponderan linealmente tres modelos distintos:

1. Modelo de trigramas: $q_{ML}(w_i|w_{i-2}, w_{i-1}) = \frac{\text{Count}(w_{i-2}, w_{i-1}, w_i)}{\text{Count}(w_{i-2}, w_{i-1})}$
2. Modelo de bigramas: $q_{ML}(w_i|w_{i-1}) = \frac{\text{Count}(w_{i-1}, w_i)}{\text{Count}(w_{i-1})}$
3. Modelo de unigramas: $q_{ML}(w_i) = \frac{\text{Count}(w_i)}{\text{Count}(\#tokens en el corpus)}$.

Cada parámetro $q(w_i|w_{i-2}, w_{i-1})$ se interpola de la siguiente manera:

$$q(w_i|w_{i-2}, w_{i-1}) = \lambda_1 \cdot q_{ML}(w_i|w_{i-2}, w_{i-1}) + \lambda_2 \cdot q_{ML}(w_i|w_{i-1}) + \lambda_3 \cdot q_{ML}(w_i)$$

Donde λ_1 , λ_2 , y λ_3 son hiper-parámetros que deben definirse manualmente y cumplir con las condiciones $\lambda_1 + \lambda_2 + \lambda_3 = 1$, y $\lambda_i \geq 0$ para todo i . Como podemos ver, con esta técnica abordamos

el problema de indefinición de probabilidades en el modelo de trigramas, puesto que si se le quiere asignar probabilidad a un trígrama no observado durante entrenamiento, este no necesariamente indefine las probabilidades pues puede recurrir a las probabilidades de los bigramas y unigramas correspondientes.

Además, se puede demostrar que el modelo interpolado define adecuadamente una distribución de probabilidad (donde definimos $V' = V \cup \{\text{STOP}\}$):

$$\begin{aligned} & \sum_{w \in V'} q(w|u, v) \\ &= \sum_{w \in V'} [\lambda_1 \cdot q_{\text{ML}}(w|u, v) + \lambda_2 \cdot q_{\text{ML}}(w|v) + \lambda_3 \cdot q_{\text{ML}}(w)] \\ &= \lambda_1 \sum_w q_{\text{ML}}(w|u, v) + \lambda_2 \sum_w q_{\text{ML}}(w|v) + \lambda_3 \sum_w q_{\text{ML}}(w) \\ &= \lambda_1 + \lambda_2 + \lambda_3 = 1 \end{aligned}$$

También es posible demostrar que $q(w|u, v) \geq 0$ para todas las palabras en V' , ya que es una suma ponderada de tres cantidades no negativas.

0.1.1 Estimación de los Valores λ

Para encontrar un valor adecuado de los hiper-parámetros λ se suele reservar una parte del conjunto de entrenamiento como datos de *validación*. Definimos $c'(w_1, w_2, w_3)$ como el número de veces que se observa el trígrama (w_1, w_2, w_3) en el conjunto de validación. Elegimos $\lambda_1, \lambda_2, \lambda_3$ para maximizar:

$$L(\lambda_1, \lambda_2, \lambda_3) = \sum_{w_1, w_2, w_3} c'(w_1, w_2, w_3) \log q(w_3|w_1, w_2)$$

sujetos a $\lambda_1 + \lambda_2 + \lambda_3 = 1$, y $\lambda_i \geq 0$ para todo i , donde

$$q(w_i|w_{i-2}, w_{i-1}) = \lambda_1 \cdot q_{\text{ML}}(w_i|w_{i-2}, w_{i-1}) + \lambda_2 \cdot q_{\text{ML}}(w_i|w_{i-1}) + \lambda_3 \cdot q_{\text{ML}}(w_i)$$

Esto generalmente se hace buscando en una grilla de posibles valores de todos los hiper-parámetros. Nótese que maximizar $L(\lambda_1, \lambda_2, \lambda_3)$ es equivalente a minimizar la perplejidad en validación.

0.2 Modelos de Descuento (Katz Back-Off)

Los modelos de descuento representan una técnica alternativa a la interpolación con el propósito de mejorar la generalización del modelo de lenguaje a datos que difieren de los utilizados en el entrenamiento. Además, estos modelos ayudan a evitar la sobreestimación de conteos para n-gramas con baja frecuencia de aparición. La idea principal detrás de los modelos de descuento es redistribuir la probabilidad de los n-gramas observados hacia aquellos que no se han observado.

Para ilustrar estos modelos, consideremos un ejemplo utilizando los conteos de bigramas que comienzan con el unígrafo “la” en un corpus, junto con sus estimaciones de máxima verosimilitud, que se presentan en la Tabla 2. Se puede notar que estas estimaciones tienden a sobrevalorar lo que se ha observado en el corpus y subestimar lo que no se ha observado.

Podemos definir los conteos “descontados” de la siguiente manera:

Frase	Conteo	$q_{ML}(w_i w_{i-1})$
la	48	
la, gata	15	15/48
la, mujer	11	11/48
la, persona	10	10/48
la, plaza	5	5/48
la, actividad	2	2/48
la, tuerca	1	1/48
la, revista	1	1/48
la, tarde	1	1/48
la, ciudad	1	1/48
la, calle	1	1/48

Cuadro 1: Ejemplo de conteos de bigramas

$$\text{Conteo}^*(x) = \text{Conteo}(x) - \beta$$

Donde β es una constante que generalmente se encuentra entre 0 y 1, siendo un valor típico 0.5. Con $\beta = 0.5$, los conteos se descuentan, es decir, se reduce la probabilidad de los datos observados, como se muestra en la Tabla 2.

Frase	Conteo	Conteo [*] (x)	$q_{BO}(w_i w_{i-1})$
la	48		
la, gata	15	14.5	14,5/48
la, mujer	11	10.5	10,5/48
la, persona	10	9.5	9,5/48
la, plaza	5	4.5	4,5/48
la, actividad	2	1.5	1,5/48
la, tuerca	1	0.5	0,5/48
la, revista	1	0.5	0,5/48
la, tarde	1	0.5	0,5/48
la, ciudad	1	0.5	0,5/48
la, calle	1	0.5	0,5/48

Cuadro 2: Ejemplo de conteos de bigramas con conteos descontados

Las nuevas estimaciones se basan en los conteos descontados, y se introduce una “masa de probabilidad faltante” como se muestra a continuación:

$$\alpha(w_{i-1}) = 1 - \sum_w \frac{\text{Conteo}^*(w_{i-1}, w)}{\text{Conteo}(w_{i-1})}$$

Por ejemplo, en nuestro caso:

$$\alpha(\text{la}) = 1 - \frac{14,5 + 10,5 + 9,5 + 4,5 + 1,5 + 0,5 + 0,5 + 0,5 + 0,5}{48} = 1 - \frac{43}{48} = \frac{5}{48}$$

La idea es redistribuir esta masa faltante entre todos los bigramas que comienzan con la palabra “la” y que no fueron observados en el corpus, como por ejemplo “rana”, “bailarina” y “puerta”.

En el modelo de Katz Back-Off de bigramas, se definen dos conjuntos: $A(w_{i-1})$ y $B(w_{i-1})$, como se detalla a continuación:

$$A(w_{i-1}) = \{w : \text{Count}(w_{i-1}, w) > 0\}$$

En nuestro ejemplo, $A(\text{la}) = \{\text{gata, mujer, persona, plaza, actividad, tuerca, revista, tarde, ciudad, calle}\}$.

Por otro lado, $B(w_{i-1})$ se define como:

$$B(w_{i-1}) = \{w : \text{Count}(w_{i-1}, w) = 0\}$$

En nuestro ejemplo, $B(\text{la}) = \{\text{rana, bailarina, puerta, ...}\}$.

Luego, la probabilidad condicional $q_{BO}(w_i|w_{i-1})$ se calcula de la siguiente manera: si la palabra w_i está en el conjunto $A(w_{i-1})$, se utiliza una estimación basada en las frecuencias relativas de la secuencia (w_{i-1}, w_i) dividida por la frecuencia de w_{i-1} . Si w_i está en el conjunto $B(w_{i-1})$, se utiliza una estimación suavizada que combina la constante $\alpha(w_{i-1})$ y las probabilidades condicionales de máxima verosimilitud de un modelo orden inferior (unigramas) $q_{ML}(w_i)$ de las palabras en el conjunto $B(w_{i-1})$. Esta última ponderación asegura que las probabilidades queden bien definidas.

$$q_{BO}(w_i|w_{i-1}) = \begin{cases} \frac{\text{Count}^*(w_{i-1}, w_i)}{\text{Count}(w_{i-1})} & \text{si } w_i \in A(w_{i-1}) \\ \frac{\alpha(w_{i-1}) q_{ML}(w_i)}{\sum_{w \in B(w_{i-1})} q_{ML}(w)} & \text{si } w_i \in B(w_{i-1}) \end{cases}$$

La constante $\alpha(w_{i-1})$ se calcula restando la suma de las frecuencias relativas de las palabras en $A(w_{i-1})$ de uno tal como se definió anteriormente.

Para extender este enfoque a trigramas, se definen conjuntos similares $A(w_{i-2}, w_{i-1})$ y $B(w_{i-2}, w_{i-1})$ para las secuencias de trigramas:

$$A(w_{i-2}, w_{i-1}) = \{w : \text{Count}(w_{i-2}, w_{i-1}, w) > 0\}$$

$$B(w_{i-2}, w_{i-1}) = \{w : \text{Count}(w_{i-2}, w_{i-1}, w) = 0\}$$

La probabilidad condicional $q_{BO}(w_i|w_{i-2}, w_{i-1})$ en un modelo de trigramas se calcula utilizando el modelo de bigrama correspondiente y aplicando la misma lógica anterior:

$$q_{BO}(w_i|w_{i-2}, w_{i-1}) = \begin{cases} \frac{\text{Count}^*(w_{i-2}, w_{i-1}, w_i)}{\text{Count}(w_{i-2}, w_{i-1})} & \text{si } w_i \in A(w_{i-2}, w_{i-1}) \\ \frac{\alpha(w_{i-2}, w_{i-1}) q_{BO}(w_i|w_{i-1})}{\sum_{w \in B(w_{i-2}, w_{i-1})} q_{BO}(w|w_{i-1})} & \text{si } w_i \in B(w_{i-2}, w_{i-1}) \end{cases}$$

Donde:

$$\alpha(w_{i-2}, w_{i-1}) = 1 - \sum_{w \in A(w_{i-2}, w_{i-1})} \frac{\text{Count}^*(w_{i-2}, w_{i-1}, w)}{\text{Count}(w_{i-2}, w_{i-1})}$$

Estos modelos de Katz Back-Off ofrecen una aproximación efectiva de las probabilidades condicionales en situaciones en las que la información disponible es limitada, al aprovechar la información de contextos más pequeños cuando no se dispone de suficientes datos para estimaciones directas.

0.3 Historia

En 1951, mientras trabajaba en los laboratorios Bell, Claude Shannon (en la Figura 1) realizó experimentos sobre la entropía del inglés, modelando el lenguaje escrito de manera estadística y predictiva [?]. Utilizando modelos de lenguaje de n-gramas, Shannon exploró la dificultad de predecir palabras basándose en las palabras anteriores.

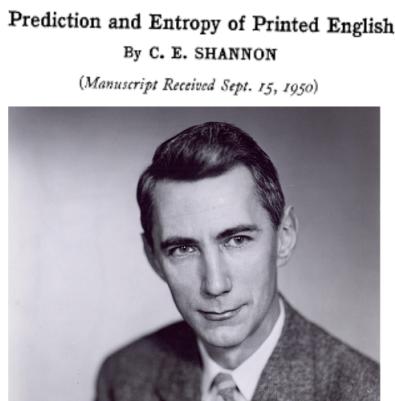


Figura 1: Imagen de Shannon.

Por otro lado, en su libro *Syntactic Structures* (1957), Noam Chomsky (en la Figura 2), un lingüista y científico cognitivo, cuestionó la capacidad de los modelos de lenguaje probabilísticos para capturar y comprender la gramática del lenguaje humano [?]. Según Chomsky, la noción de “gramaticalmente correcto” no puede ser equiparada a “significativo” en un sentido probabilista. Para ilustrar esto, presentó dos oraciones ficticias, ambas carentes de sentido:

1. Colorless green ideas sleep furiously.
2. Furiously sleep ideas green colorless.

Aunque ambas oraciones carecen de significado, Chomsky argumentó que solo la primera se considera gramaticalmente correcta por los hablantes de inglés. Además, enfatizó que la corección gramatical en inglés no puede determinarse únicamente mediante aproximaciones estadísticas. Aunque es poco probable que ninguna de las dos oraciones (1) o (2) haya surgido en documentos escritos en inglés, un modelo estadístico como los modelos de lenguaje vistos en este capítulo las consideraría igualmente “remotas” en relación al inglés. Sin embargo, la oración (1) es gramaticalmente correcta, mientras que la oración (2) no lo es, lo que destaca las limitaciones de los enfoques estadísticos para capturar la gramática. Estos argumentos retrasaron el estudio de los modelos de lenguaje probabilísticos durante varios años [?].

0.4 Conclusiones

El cálculo de probabilidades en modelos de lenguaje probabilísticos implica tres pasos:

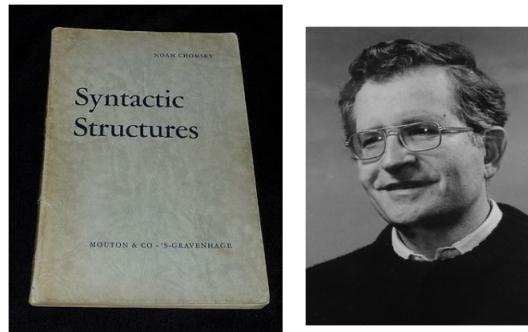


Figura 2: Imagen de Chomsky

1. Expandir $p(w_1, w_2, \dots, w_n)$ usando la regla de la Cadena.
2. Aplicar los supuestos Independencia de Markov
$$p(w_i|w_1, w_2, \dots, w_{i-2}, w_{i-1}) = p(w_i|w_{i-2}, w_{i-1}).$$
3. Suavizar las estimaciones utilizando conteos de orden inferior.

No obstante, los modelos de lenguaje de bigramas o trigramas (que consideran dos palabras anteriores como contexto) tienen limitaciones en contextos largos y no pueden aprovechar contextos similares. Por ejemplo, consideremos los contextos:

- c_1 : Despu  s de comer cereales
- c_2 : Luego de desayunar avena

Aunque esperar  mos que las distribuciones de probabilidad $p(w|c_1)$ y $p(w|c_2)$ fueran similares, dado que c_1 y c_2 casi no comparten palabras, los modelos de n-gramas que se limitan a contar frecuencia de palabras no pueden capturar estas similitudes entre contextos.

Otros m  todos para mejorar los modelos de lenguaje incluyen introducir variables latentes para representar t  picos, conocidos como modelos de t  picos [?] presentados en el Cap  tulo ?? . O alternativamente, reemplazar $p(w_i|w_1, w_2, \dots, w_{i-2}, w_{i-1})$ con una red neuronal predictiva y una “capa de embedding” para representar mejor contextos m  s grandes y aprovechar similitudes entre palabras en el contexto. [?]

Los modelos de lenguaje modernos utilizan redes neuronales profundas en su estructura principal y tienen un vasto espacio de par  metros como se ver   en el Cap  tulo ?? .