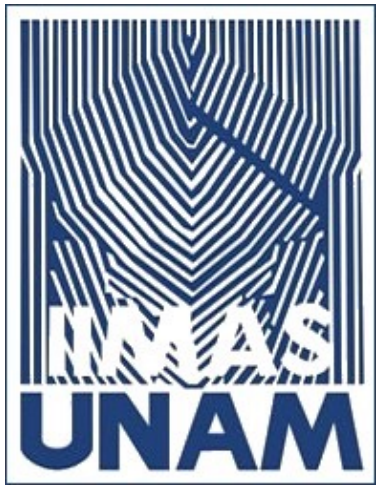
Procesamiento de Lenguaje Natural

Autocompletado y Modelos del Lenguaje



Dra. Helena Gómez Adorno
helena.gomez@iimas.unam.mx

Dra. Gemma Bel
gbele@iingen.unam.mx

Correo del curso:
pln.cienciadedatos@gmail.com

Asistente:
Luis Ramon Casillas

Esta semana veremos:

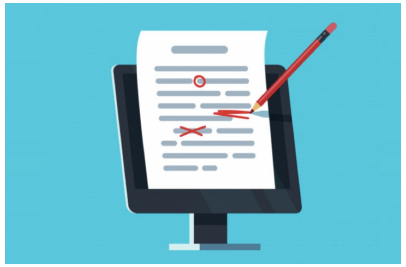
- Creación de Modelos del lenguaje a partir de un corpus de texto
  - Estimación de probabilidades de secuencias de palabras
  - Estimación de probabilidades de una palabras siguiendo a una secuencia de palabras
- Aplicación del modelo para autocompletar una oración

Otras aplicaciones

- Reconocimiento de voz

P(El es conde) > P(El esconde)

- Corrección gramatical



Ay una tetera preparada para la convención de enfermería

P(Ay una tetera preparada) > P (Hay una tetera preparada)

Contenido

- Procesar un corpus de texto a un modelo del lenguaje basado en n-grama
- Palabras fuera de vocabulario
- Suavizado para n-gramas no vistos previamente
- Evaluación del modelo del lenguaje

Autocompletado de oraciones

N-grama

Un n-grama es una secuencia de N palabras

Corpus: Yo soy feliz porque estoy aprendiendo

Unigramas:{Yo, soy, feliz, porque, estoy, aprendiendo}

Bigramas: {Yo soy, soy feliz, feliz porque, porque estoy, …}

Trigramas: {Yo soy feliz, soy feliz porque, feliz porque estoy, …}

Notación de secuencia

Corpus: This is great ... teacher drinks tea.

$w_1 \quad w_2 \quad w_3 \qquad\qquad w_{498} \quad w_{499} \quad w_{500}$

$$m = 500$$

$$w_1^m = w_1 w_2 w_3 \ldots . w_m$$

$$w_1^3 = w_1 w_2 w_3$$

$$w_{m-2}^m = w_{m-2} w_{m-1} w_m$$

Probabilidad de unigramas

Dado el siguiente corpus: *I am happy because I am learning.*

Tamaño del corpus m=7

$$P(I) = \frac{2}{7}$$

$$P(happy) = \frac{1}{7}$$

Para generalizar la probabilidad de un unigrama es $P(w) = \frac{C(w)}{m}$

Probabilidad de bigramas

Corpus: I am happy because I am learning

$$P(am|I) = \frac{C(I\ am)}{C(I)} = \frac{2}{2} = 1$$

$$P(happy|I) = \frac{C(I\ happy)}{C(I)} = \frac{0}{2} = 0$$

✖ I happy

$$P(learning|am) = \frac{C(am\ learning)}{C(am)} = \frac{1}{2}$$

La probabilidad de un bigrama es $$P(y|x) = \frac{C(x\ y)}{\sum_w C(x\ w)} = \frac{C(x\ y)}{C(x)}$$

Probabilidad de trigramas

Dado el siguiente corpus: *I am happy because I am learning.*

$$P(happy|I\ am) = \frac{C(I\ am\ happy)}{C(I\ am)} = \frac{1}{2}$$

La probabilidad de un trigrama es $P\left(w_3 \mid w_1^2\right) = \frac{C\left(w_1^2 w_3\right)}{C\left(w_1^2\right)}$

$$C\left(w_1^2 w_3\right) = C\left(w_1 w_2 w_3\right) = C\left(w_1^3\right)$$

Probabilidad de n-grama

$$P\left(w_N \mid w_1^{N-1}\right) = \frac{C\left(w_1^{N-1} w_N\right)}{C\left(w_1^{N-1}\right)}$$

$$C\left(w_1^{N-1} w_N\right) = C\left(w_1^N\right)$$

Probabilidad de una secuencia

- Dada una oración, cual es su probabilidad?

$$P(the\ teacher\ drinks\ tea) = ?$$

- Probabilidad condicional y regla de la cadena

$$P(B|A) = \frac{P(A,B)}{P(A)} \Rightarrow P(A,B) = P(A)P(B|A)$$

$$P(A,B,C,D) = P(A)P(B|A)P(C|A,B)P(D|A,B,C)$$

Probabilidad de una secuencia

- Dada una oración, cual es su probabilidad?

$$P(the\ teacher\ drinks\ tea) =$$

$$P(the)\ P(teacher|the)P(drinks|the\ teacher)$$

$$P(tea|the\ teacher\ drinks)$$

Cuando la oración no existe en el corpus

- Uno de los principales problemas al calcular las probabilidades anteriores es que el corpus rara vez contiene exactamente las mismas frases en las que calculó sus probabilidades.
- Por lo tanto, puede terminar obteniendo fácilmente una probabilidad de 0.

$$P(tea|the\ teacher\ drinks) = \frac{C(the\ teacher\ drinks\ tea)}{C(the\ teacher\ drinks)}$$

Probablemente 0

Cuando la oración no existe en el corpus

- La suposición de Markov indica que solo importa la última palabra.

Corpus: *the teacher drinks tea*

$$P(tea|the\ teacher\ drinks) \approx P(tea|drinks)$$

$P(teacher|the)$
$P(drinks|teacher)$
$P(tea|drinks)$

$$P(the\ teacher\ drinks\ tea) =$$

$$P(the)P(teacher|the)P(drinks|the\ teacher)P(tea|the\ teacher\ drinks)$$

$$P(the)P(teacher|the)P(drinks|teacher)P(tea|drinks)$$

Cuando la oración no existe en el corpus

- La suposición de Markov indica que solo importa la última palabra.

$$\text{Bigram} \quad P\left(w_n \mid w_1^{n-1}\right) \approx P\left(w_n \mid w_{n-1}\right)$$

$$\text{N-gram} \quad P\left(w_n \mid w_1^{n-1}\right) \approx P\left(w_n \mid w_{n-N+1}^{n-1}\right)$$

- Modelado de una oración completa:

$$P\left(w_1^n\right) \approx \prod_{i=1}^{n} P\left(w_i \mid w_{i-1}\right)$$

$$P\left(w_1^n\right) \approx P\left(w_1\right) P\left(w_2 \mid w_1\right) \ldots P\left(w_n \mid w_{n-1}\right)$$

Símbolos de inicio y fin de oración

- Por lo general, comenzamos y terminamos una oración con los siguientes tokens respectivamente: <s> </s>.

- Al calcular probabilidades usando un unigrama, puede agregar un <s> al principio de la oración.
- Para el token de final de oración </s>, solo necesita uno, incluso si es un N-gram.

Símbolos de inicio y fin de oración

the teacher drinks tea

$$P(the\ teacher\ drinks\ tea) \approx P(the)P(teacher|the)P(drinks|teacher)P(tea|drinks)$$

<s> the teacher drinks tea

$$P(<s>\ the\ teacher\ drinks\ tea) \approx P(the|<s>)P(teacher|the)P(drinks|teacher)P(tea|drinks)$$

Símbolos de inicio y fin de oración

- Trigramas

$$P(the\ teacher\ drinks\ tea) \approx P(the)P(teacher|the)P(drinks|the\ teacher)P(tea|teacher\ drinks)$$

the teacher drinks tea => <s> <s> the teacher drinks tea

$$P(w_1^n) \approx P(w_1|<s>\ <s>)P(w_2|<s>\ w_1)...P(w_n|w_{n-2}\ w_{n-1})$$

Para generalizar a un modelo de lenguaje basado en N-gramas, puede agregar N-1 tokens de inicio <s>.

Símbolos de fin de oración </s>

$$P(y|x) = \frac{C(x\ y)}{\sum_w C(x\ w)} = \frac{C(x\ y)}{C(x)}$$

Corpus:

<s> Lyn drinks chocolate

<s> John drinks

$$\sum_w C(drinks\ w) = 1$$

$$C(drinks) = 2$$

Símbolos de fin de oración </s>

Corpus

<s> yes no

<s> yes yes

<s> no no

Sentences of length 2:

<s> yes yes

<s> yes no

<s> no no

<s> no yes

$$P(< s > \text{ yes yes}) =$$

$$P(\text{yes} \mid < s >) \times P(\text{yes} \mid \text{yes}) =$$

$$\frac{C(< s > \text{ yes})}{\sum_w C(< s > w)} \times \frac{C(\text{yes yes})}{\sum_w C(\text{yes } w)} =$$

$$\frac{2}{3} \times \frac{1}{2} = \frac{1}{3}$$

Símbolos de fin de oración </s>

Corpus

<s> yes no

<s> yes yes

<s> no no

Sentences of length 2:

<s> yes yes

<s> yes no

<s> no no

<s> no yes

$$P(< s > \text{ yes yes}) = \frac{1}{3}$$

$$P(< s > \text{ yes no}) = \frac{1}{3}$$

$$P(< s > \text{ no no}) = \frac{1}{3}$$

$$P(< s > \text{ no yes}) = 0$$

$$\sum_{\text{2 word}} P(\cdots) = 1$$

Símbolos de fin de oración </s>

Corpus

<s> yes no

<s> yes yes

<s> no no

Sentences of length 3:

<s> yes yes yes

<s> yes yes no

...

<s> no no no

$$P(< s > \text{ yes yes yes}) = \dots$$

$$P(< s > \text{ yes yes no}) = \dots$$

$$\dots \qquad = \dots$$

$$P(< s > \text{ no no no}) = \dots$$

$$\sum_{3 \text{ word}} P(\cdots) = 1$$

Símbolos de fin de oración </s>

Corpus

<s> yes no

<s> yes yes

<s> no no

$$\sum_{\text{2 word}} P(\cdots) \qquad + \sum_{\text{3 word}} P(\cdots) \qquad + \ldots = 1$$

Símbolos de fin de oración </s>

- Bigramas

<s> the teacher drinks tea => <s> the teacher drinks tea </s>

$$P(the|<s>)P(teacher|the)P(drinks|teacher)P(tea|drinks)P(</s>|tea)$$

Corpus:
<s> Lyn drinks chocolate </s>
<s> John drinks </s>

$$\sum_w C(drinks\ w) = 2$$

$$C(drinks) = 2$$

Ejemplo - Bigramas

Corpus

<s> Lyn drinks chocolate </s>

<s> John drinks tea </s>

<s> Lyn eats chocolate </s>

$$P(sentence) =$$

$$P(John|<s>) = \frac{1}{3}$$

$$P(chocolate|eats) = \frac{1}{2}$$

$$P(</s>|tea) = \frac{1}{1}$$

$$P(Lyn|<s>) = ?$$

Modelo del lenguaje basado en n-gramas

Conceptos

- Matriz de conteo
- Matriz de probabilidad
- Modelo de lenguaje
- Logaritmo de la probabilidad para evitar subdesbordamiento
- Modelo de lenguaje generativo

Matriz de conteo

- Las filas corresponden a los N-1-gramas del corpus.
- Las columnas corresponden a las palabras únicas del corpus.
- A continuación se muestra un ejemplo de la matriz de recuento de un bigrama.

- Bigram count matrix

Corpus: <s>I study I learn</s>

	<s>	</s>	I	study	learn
<s>	0	0	1	0	0
</s>	0	0	0	0	0
I	0	0	0	1	1
study	0	0	1	0	0
learn	0	1	0	0	0

"study I" bigram

Matriz de probabilidades

- Se divide cada celda por la suma de la fila. Se puede usar la siguiente formula:

$$P\left(w_n \mid w_{n-N+1}^{n-1}\right) = \frac{C\left(w_{n-N+1}^{n-1}, w_n\right)}{C\left(w_{n-N+1}^{n-1}\right)}$$

$$\text{sum}(row) = \sum_{w \in V} C\left(w_{n-N+1}^{n-1}, w\right) = C\left(w_{n-N+1}^{n-1}\right)$$

- Bigram count matrix

"study I" bigram

Corpus: <s>I study I learn</s>

	<s>	</s>	I	study	learn
<s>	0	0	1	0	0
</s>	0	0	0	0	0
I	0	0	0	1	1
study	0	0	1	0	0
learn	0	1	0	0	0

Probability matrix

	<s>	</s>	I	study	learn
<s>	0	0	1	0	0
</s>	0	0	0	0	0
I	0	0	0	0.5	0.5
study	0	0	1	0	0
learn	0	1	0	0	0

Modelo del lenguaje

- Ahora, dada la matriz de probabilidad, puede generar el modelo de lenguaje.
- Se puede calcular la probabilidad de la oración y la predicción de la siguiente palabra. Para calcular la probabilidad de una secuencia, necesitaba calcular:

$$P\left(w_1^n\right) \approx \prod_{i=1}^{n} P\left(w_i \mid w_{i-1}\right)$$

- Para evitar el desbordamiento, puede multiplicar por el logaritmo de la probabilidad.

$$\log\left(P\left(w_1^n\right)\right) \approx \sum_{i=1}^{n} \log\left(P\left(w_i \mid w_{i-1}\right)\right)$$

Modelo del lenguaje generativo

Corpus:

<s> Lyn drinks chocolate </s>

<s> John drinks tea </s>

<s> Lyn eats chocolate </s>

1. (<s>, Lyn) or (<s>, John)?
2. (Lyn,eats) or (Lyn,drinks) ?
3. (drinks,tea) or (drinks,chocolate)?
4. (tea,</s>) - always

- Algoritmo:
- Elegir inicio de la oración
- Elegir el siguiente bigrama empezando con la palabra previa
- Continuar hasta que </s> sea elegido

Evaluación del modelo del lenguaje

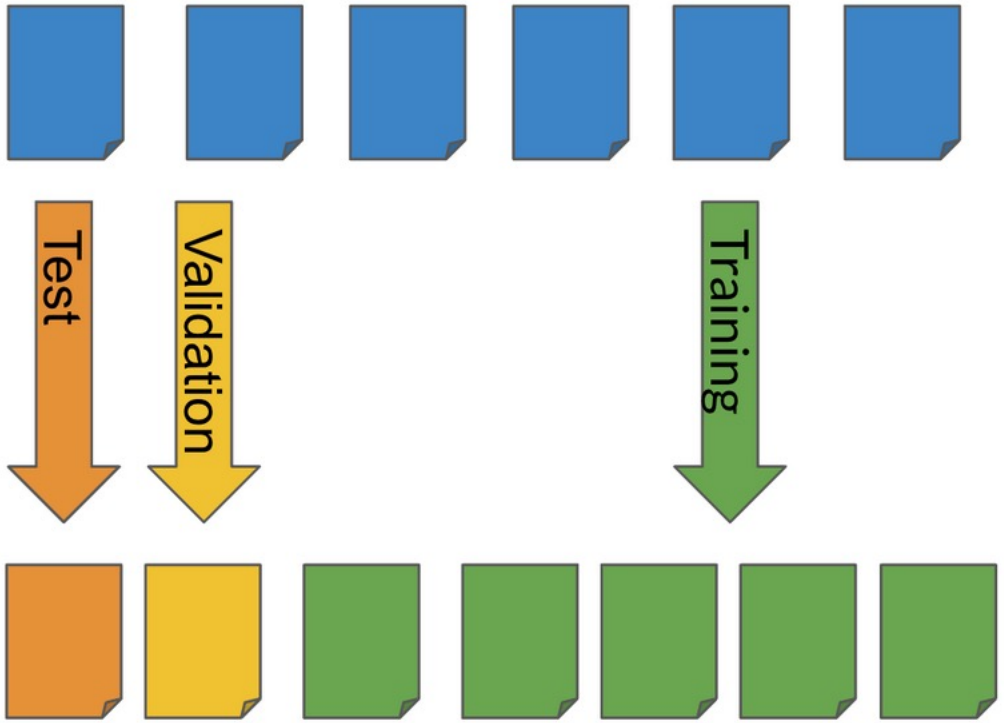
- Discutiremos las divisiones de corpus para evaluación y la medida de perplejidad.

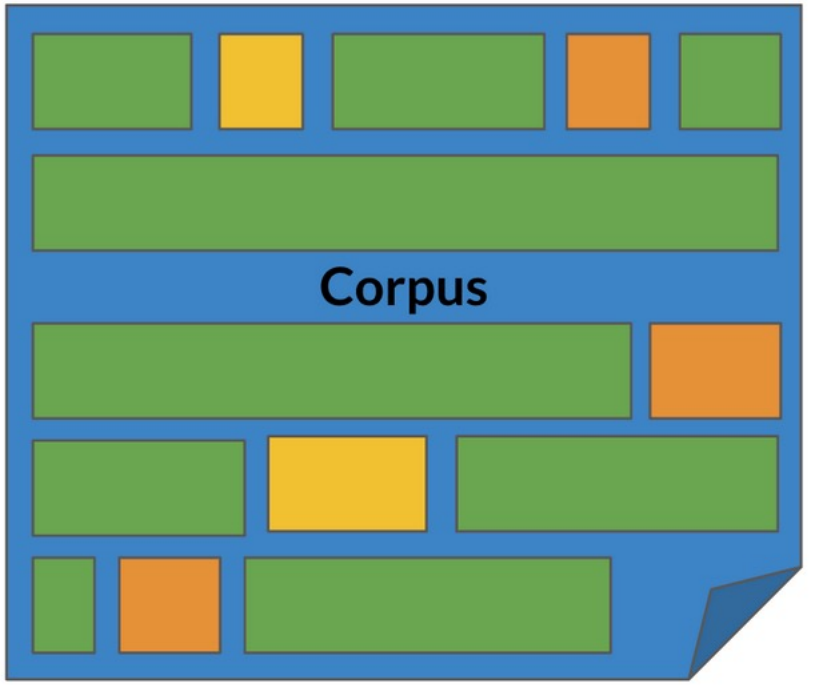
Divisiones de Entrenamiento / Validación / Prueba

- Corpus más pequeños:
- 80% entrenamiento
- 10% validación
- 10% prueba

- Corpus más grande:
- 98% entrenamiento
- 1% validación
- 1% prueba

Dos métodos para la división de datos





Perplejidad (Perplexity)

- La perplejidad se usa para decirnos si un conjunto de oraciones parece que fueron escritas por humanos en lugar de por un programa simple que elige palabras al azar. Es más probable que un texto escrito por humanos tenga una menor perplejidad, mientras que un texto generado por la elección de palabras al azar tendría una mayor perplejidad.

$$PP(W) = P s_1, s_2, \ldots, s_m^{-\frac{1}{m}}$$

W → conjunto de prueba que contiene *m* oraciones *s*

S_i → i-ésima oración en el conjunto de prueba, cada una termina con </s>

m →número de todas las palabras en el conjunto de prueba W, incluyendo </s> pero no <s>

Perplejidad (Perplexity)

E.g. m=100

$$P(W) = 0.9 => PP(W) = 0.9^{-\frac{1}{100}} = 1.00105416$$

$$P(W) = 10^{-250} => PP(W) = (10^{-250})^{-\frac{1}{100}} \approx 316$$

Perplejidad pequeña = mejor modelo

PP Modelos a nivel de carácter < PP modelos a nivel de palabra

Perplejidad (Perplexity)

$$PP(W) = \sqrt[m]{\prod_{i=1}^{m} \prod_{j=1}^{s_i} \frac{1}{Pw_j^{(i)} \mid w_{j-1}^{(i)}}}$$

$w_j^{(i)}$ →j corresponde a la j-ésima palabra de la i-ésima oración. Si tuviera que concatenar todas las oraciones, entonces w_i es la i-ésima palabra en el conjunto de prueba. Para calcular el logaritmo de la perplejidad, se pasa de:

$$PP(W) = \sqrt[m]{\prod_{i=1}^{m} \frac{1}{Pw_i \mid w_{i-1}}} \quad \Longrightarrow \quad \log PP(W) = -\frac{1}{m} \sum_{i=1}^{m} \log_2 \left(P\left(w_i \mid w_{i-1}\right)\right)$$

Ejemplos

Training 38 million words, test 1.5 million words, WSJ corpus
Perplexity Unigram: 962 Bigram: 170 Trigram: 109

Unigram

Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives

Bigram

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

Trigram

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

[Figure from *Speech and Language Processing* by Dan Jurafsky et. al]

Palabras fuera de vocabulario

- Muchas veces, se encontrará con palabras desconocidas en el corpus. Entonces, ¿cómo eliges tu vocabulario? ¿Qué es un vocabulario?

- Un vocabulario es un conjunto de palabras únicas respaldadas por su modelo de lenguaje. En algunas tareas como el reconocimiento de voz o la respuesta a preguntas, encontrará y generará palabras solo a partir de un conjunto fijo de palabras. De ahí un **vocabulario cerrado**.

- **Vocabulario abierto** significa que puede encontrar palabras que no pertenecen al vocabulario, como el nombre de una nueva ciudad en el conjunto de formación. Discutiremos una idea que nos permitiría manejar palabras desconocidas.

Uso del tag <UNK> en el corpus

- Crear vocabulario **V**
- Reemplazar cualquier palabra en el corpus y no en V por <UNK>
- Contar las probabilidades con <UNK> como con cualquier otra palabra

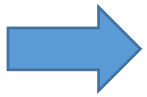
Ejemplo

Corpus

<s> Lyn drinks chocolate </s>

<s> John drinks tea </s>

<s> Lyn eats chocolate </s>



Corpus

<s> Lyn drinks chocolate </s>

<s> <UNK> drinks <UNK> </s>

<s> Lyn <UNK> chocolate </s>

Frecuencia mínima f=2

Vocabulary

Lyn, drinks, chocolate

Input query

<s>Adam drinks chocolate</s>

<s><UNK> drinks chocolate</s>

Cómo crear el vocabulario V

- Frecuencia mínima de palabras f
- Max |V|, incluye palabras por frecuencia
- Use <UNK> escasamente (Porqué?)
- Perplejidad: solo compare LM con la misma V

N-gramas faltantes en el corpus de entrenamiento

- Problema: N-gramas compuestos de palabras conocidas pueden estar faltantes en el conjunto de entrenamiento:
- Las palabras “John”, “eats” están en corpus, pero el n-grama “John eats” no está.
- Su conteo no puede ser usar para la estimación de probabilidad.

$$P(w_n|w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1}, w_n)}{C(w_{n-N+1}^{n-1})}$$

Puede ser 0

Trataremos tres conceptos principales para lidiar con n-gramas que faltan, el suavizado y el retroceso e interpolación.

N-gramas faltantes en el corpus de entrenamiento

- Suavizado agregando 1 (Add 1, Laplacian)

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1},w_n)+1}{\sum_{w\in V}(C(w_{n-1},w)+1)} = \frac{C(w_{n-1},w_n)+1}{C(w_{n-1})+V}$$

- Suavizado agregando k (Add 1, Laplacian)

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1},w_n)+k}{\sum_{w\in V}(C(w_{n-1},w)+k)} = \frac{C(w_{n-1},w_n)+k}{C(w_{n-1})+k*V}$$

Retroceso

- Si falta N-grama => usar (N-1) -grama,…: Usar el nivel inferior de N-gramas (es decir, (N-1)-grama, (N-2)-grama, hasta unigrama) distorsiona la probabilidad distribución. Especialmente para corpus más pequeños, es necesario descontar alguna probabilidad de los N-gramas de nivel superior para usarla para N-gramas de nivel inferior.
- Descuento de probabilidad, p. Ej. Retroceso de Katz: hace uso de descuentos.
- Retroceso “estúpido”: si falta la probabilidad de N-grama de orden superior, se utiliza la probabilidad de N-grama de orden inferior, simplemente multiplicada por una constante. Se demostró experimentalmente que una constante de aproximadamente 0,4 funcionaba bien.

Ejemplo

Corpus

<s> Lyn drinks chocolate </s>

<s> John drinks tea </s>

<s> Lyn eats chocolate </s>

$$P(chocolate|John\ drinks) = ?$$

$$0.4 \times P(chocolate|drinks)$$

Interpolación

$$\hat{P}w_n \mid w_{n-2}w_{n-1} = \lambda_1 \times Pw_n \mid w_{n-2}w_{n-1} + \lambda_2 \times Pw_n \mid w_{n-1} + \lambda_3 \times Pw_n$$

Where

$$\sum_i \lambda_i = 1$$