# Modelado de Lenguaje: Notas Complementarias

Franco M. Luque

**Resumen**

Este documento contiene notas complementarias a las notas de Michael Collins.

## 1. Evaluación

Sean $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ las oraciones del corpus de evaluación.

La log-likelihood es la log-probabilidad del corpus de evaluación (a más alta mejor):

$$L = \sum_{i=1}^{m} \log_2(p(x^{(i)})).$$

La cross-entropy es el opuesto de la log-probabilidad promedio de cada token del corpus de evaluación:

$$H = \frac{-1}{M} * L$$

a donde $M$ es la cantidad total de tokens. Se puede interpretar como la cantidad de bits promedio necesaria para codificar la predicción del siguiente token (a más baja mejor).

La perplexity es:

$$2^H.$$

Se puede interpretar como la cantidad de opciones promedio que el modelo tiene para predecir el siguiente token (a más baja mejor).

## 2. Interpolación

Forma general de la interpolación para $n$-gramas:

$$q(x_n|x_1 \ldots x_{n-1}) = \sum_{i=1}^{n} \lambda_i(x_i \ldots x_{n-1})\, q_{ML}(x_n|x_i \ldots x_{n-1})$$

a dónde, para $i \leq n-1$,

$$\lambda_i(x_i \ldots x_{n-1}) = \left(1 - \sum_{j=1}^{i-1} \lambda_j\right) \frac{c(x_i \ldots x_{n-1})}{c(x_i \ldots x_{n-1}) + \gamma}$$

y

$$\lambda_n() = 1 - \sum_{j=1}^{n-1} \lambda_j.$$

El hiperparámetro $\gamma$ indica la intensidad del suavizado.

# 3. Back-Off con Discounting

Forma general del back-off con discounting para $n$-gramas:

$$q_D(x_i|x_1 \dots x_{i-1}) = \begin{cases} \frac{c^*(x_1 \dots x_i)}{c(x_1 \dots x_{i-1})} & \text{if } x_i \in \mathcal{A}(x_1 \dots x_{i-1}) \\ \alpha(x_1 \dots x_{i-1}) \frac{q_D(x_i|x_2 \dots x_{i-1})}{denom(x_1 \dots x_{i-1})} & \text{if } x_i \in \mathcal{B}(x_1 \dots x_{i-1}) \end{cases}$$

para $i > 1$, y

$$q_D(x_1) = \frac{c(x_1)}{c()}.$$

Acá,

$$c^*(x_1 \dots x_i) = c(x_1 \dots x_i) - \beta$$

a donde $\beta$ es un hiperparámetro que indica el valor de descuento.

## 3.1. Cálculo de $\alpha$

Se puede calcular directamente como lo dice su definición:

$$\alpha(x_1 \dots x_i) = 1 - \sum_{x \in \mathcal{A}(x_1 \dots x_i)} \frac{c^*(x_1 \dots x_i x)}{c(x_1 \dots x_i)}$$

También se puede demostrar que

$$\alpha(x_1 \dots x_i) = \begin{cases} \frac{\beta |\mathcal{A}(x_1 \dots x_i)|}{c(x_1 \dots x_i)} & \text{if } \mathcal{A}(x_1 \dots x_i) \neq \emptyset \\ 1 & \text{otherwise.} \end{cases}$$

y por lo tanto se puede calcular de esta manera.

## 3.2. Cálculo del Denominador Normalizador

Se puede calcular usando la siguiente equivalencia

$$\begin{aligned} denom(x_1 \dots x_{i-1}) &= \sum_{x \in \mathcal{B}(x_1 \dots x_{i-1})} q_D(x|x_2 \dots x_{i-1}) \\ &= 1 - \sum_{x \in \mathcal{A}(x_1 \dots x_{i-1})} q_D(x|x_2 \dots x_{i-1}) \end{aligned}$$

ya que el conjunto $\mathcal{A}$ se conoce.

Como $x \in \mathcal{A}(x_1 \dots x_{i-1})$ implica que $x \in \mathcal{A}(x_2 \dots x_{i-1})$, entonces se aplica el primer caso en la definición de $q_D$, quedando

$$\begin{aligned} denom(x_1 \dots x_{i-1}) &= \sum_{x \in \mathcal{B}(x_1 \dots x_{i-1})} q_D(x|x_2 \dots x_{i-1}) \\ &= 1 - \sum_{x \in \mathcal{A}(x_1 \dots x_{i-1})} q_D(x|x_2 \dots x_{i-1}) \\ &= 1 - \sum_{x \in \mathcal{A}(x_1 \dots x_{i-1})} \frac{c^*(x_2 \dots x_{i-1} x)}{c(x_2 \dots x_{i-1})} \end{aligned}$$