N-gramas

Los N-gramas son secuencias de palabras que se sacaron de un texto, con el propósito de predecir la probabilidad de que una palabra siga en una secuencia dada. Todo esto basándose en las palabras anteriores. Uno de los más grandes problemas es la falta de datos, que ocasiona la probabilidad cero. Para evitar eso se ocupan las siguientes técnicas:

Absolut discounting

A.D. es una técnica de suavizado que ajusta las probabilidades de los N-gramas restando una cantidad fija de las frecuencias. Con la probabilidad ahorrada se vuelve a repartir a los N-gramas no observados para evitar que tengan probabilidad cero. La fórmula matemática es la siguiente:

$$P_{AD}(w_n|w_{n-1},\dots w_{N+1}) = \begin{cases} \dfrac{C(w_{n-N+1},\dots,w_n)-D}{C(w_{n-N+1},\dots,w_{n-1})} & ; \text{ si } C(w_{n-N+1},\dots,w_n)>0 \\ \lambda(w_{n-1},\dots,w_{n-N+2})\cdot P_{cont}(w_n) & ; \text{ si } C(w_{n-N},\dots,w_n)=0 \end{cases}$$

Ejemplo:

Imaginemos un modelo de biagramas, donde en el corpus aparece "gato negro" 3 veces y "gato blanco" 1 vez. Sin embargo, "gato café" nunca aparece. En este caso, Absolut Discount le daría parte de la probabilidad de "gato negro" y "gato blanco a "gato cafe" para que así no tenga probabilidad cero.

Kneser-Ney Smoothing

Esta es una técnica más avanzada de suavizado que considera cuántas veces ocurre un N-grama, así como cuántos contextos diferentes existen antes de una palabra. Esta técnica sirve mucho para cuando el corpus tiene palabras extrañas, ya que este se basa en la probabilidad de que una palabra aparezca en un contexto nuevo. Como tal, Kneser-Ney resuelve problemas donde algunas palabras tienen muchas ocurrencias en un contexto, pero poca en otro. La fórmula básica matemática es:

$$P_{KN}(w_n|w_{n-1},\dots,w_{n-N+1}) = \frac{\max(C(w_{n-N+1},\dots,w_n)-D,0)}{C(w_{n-N+1},\dots,w_{n-1})}$$

La prob. de continuación es importante ya que está basada en cuántos contextos preceden a w_n, lo que nos dice que una palabra que es común en muchos contextos, pero rara en el corpus total, tendrá una mais alta en este modelo.

Ejemplo:

Si la palabra "reina" aparece mucho después de "la", pero pocas veces en otras situaciones, KN ajustará la probabilidad de que "reina" aparezca en contextos nuevos. Por otro lado, la palabra "hombre", que puede aparecer en muchos contextos diferentes, tendrá una probabilidad mayor en caso de nuevos contextos.