Absolute Discounting.

El problema que resuelve es que si un n-grama no es visto su probabilidad sea 0. Lo cual no permite tener un modelo generalizado

Mediante esta técnica, se descuenta una probabilidad a los n-gramas observados y se redistribuye entre los n-gramas no observados

La expresión matemática es la siguiente:

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

Donde:

- $C(w_{n-1}, w_n)$ es la frecuencia del N-grama
- D es el descuento
- $\lambda(w_{n-1})$ es un factor de normalización que asegura que la suma de probabilidades sea 1
- $P_{AD}(w_n)$ es una probabilidad de respaldo

Por ejemplo

Si en un corpus se tienen los bigramas "the dog" (frecuencia 5) y "the cat" (frecuencia 3), se puede aplicar un descuento $D = 0.75$, la probabilidad de "the dog" será ajustada y esa probabilidad será asignada a bigramas no observados como "the bird"

Kneser-Key smoothing

Este modelo considera no solo la frecuencia de los N-gramas sino también el contexto en el que aparece.

Se basa en una variación del descuento absoluto y reestima la probabilidad de respaldo considerando cuantos contextos diferentes preceden a una palabra. Esto ayuda a asignar una mayor probabilidad a palabras que aparecen en muchos contextos diferentes.

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

Donde P_{KN} se define en función del número de N gramas únicos que terminan con w_n

$$P_{KN}(w_n) = \frac{\text{count of unique contexts with } w_n}{\text{total number of unique bigram contexts}}$$

Por ejemplo, la probabilidad de la palabra "the" se ajustará dependiendo de la cantidad de sustantivos con los que aparezca.