

Absolute discounting

ES una técnica de suavizado en la que se resta una cantidad constante a las frecuencias de los N-gramas observados para redistribuir esta probabilidad a los N-gramas no observados.

Problemática que resuelve:

cuando los N-gramas tienen frecuencia de 0 en el corpus de entrenamiento esta técnica nos ayuda a asignar una mayor probabilidad a aquellos no observados sin sesgar el modelo.

Expresión matemática:

Sea $P(w_i | w_{i-n+1}^{i-1})$ la probabilidad de un N-grama dado sus N-1 palabras previas.

$$\frac{\text{Count}(w_{i-n+1}^i) - D}{\text{Count}(w_{i-n+1}^{i-1})} \quad \text{si } \text{Count}(w_{i-n+1}^i) > 0$$

$$\lambda(w_{i-n+1}^{i-1}) P_{\text{backoff}}(w_i) \quad \text{en otro caso}$$

D es el parámetro de descuento, P_{backoff} es la prob. asignada al N-grama de menor orden. $\lambda(w_{i-n+1}^{i-1})$ es un factor de normalización para que todo sume 1.

Ejemplo: si en el corpus, la frase "el perro come" tiene una f. de 3 y la de "el gato duerme" no aparece, se distribuye un poco de prob. a "el gato duerme".

Kneser-Ney

Es más sofisticado pero este, además tiene en cuenta la diversidad de contextos en los que una palabra ocurre, proporcionando unos mejores prob. a aquellos de menor orden.

Expresión:

$$P_{KN}(w_i | w_{i-n+1}^{i-1}) = \frac{\max(\text{count}(w_{i-n+1}^{i-1}) - D, 0)}{\text{count}(w_{i-n+1}^{i-1})} + \lambda(w_{i-n+1}^{i-1}) P_{KN}(w_i | w_{i-n+2}^{i-1})$$

D es el parámetro de descuento $\lambda(w_{i-n+1}^{i-1})$
 ~~$P_{KN}(w_i | w_{i-n+2}^{i-1})$~~ es un factor de ajuste para las prob. que asegura que las prob sumen 1.

$P_{KN}(w_i | w_{i-n+2}^{i-1})$ es la prob de respaldo (backoff) de menor orden.

Ejemplo:

Si la palabra "come" aparece en muchas frases diff. El suavizado reduce su prob en el modelo, porque aparecer en múltiples contextos no la hace especialmente predecible en uno.