Luis Maximiliano López Ramírez A00833321

Absolute Discounting y Smoothing Knesser-Ney para n-gramas

Las técnicas de Absolute Discounting y Knesser-Ney Smoothing son métodos de suavizado utilizados en el procesamiento de lenguaje natural, particularmente para modelar n-gramas. A continuación, se presenta un análisis de cada técnica:

1. Absolute Discounting

El Absolute Discounting es una técnica de suavizado que ajusta las probabilidades de las n-gramas observadas restando un valor fijo d, de sus frecuencias. Esta técnica intenta corregir el problema de sobreajuste en n-gramas de baja frecuencia, asignando parte de la probabilidad total a n-gramas no observados.

- Definición
  La probabilidad suavizada de un n-grama de longitud n, dado su prefijo de longitud $n-1$, se define como:

$$P(w_n | w_1, \ldots, w_{n-1}) = \max\left(\frac{c(w_1, \ldots, w_n) - d, 0)}{c(w_1, \ldots, w_{n-1})}\right) \rightarrow$$

$$\rightarrow + \lambda P_{backoff}(w_n | w_2, \ldots, w_{n-1}),$$

donde:

- $c(w_1, \ldots, w_n)$: Conteo del n-grama $w_1, \ldots, w_n$.
- $c(w_1, \ldots, w_{n-1})$: Conteo del prefijo $w_1, \ldots, w_{n-1}$
- δ: Descuento absoluto, un valor fijo que se resta de los conteos
- λ: Un parámetro de normalización que asegura que la distribución de probabilidad total suma 1.
- $P_{katz}(w_n | w_1, \ldots, w_{n-1})$: Modelo de respaldo que estima la probabilidad usando n-gramas de menor orden (en este caso, n-1 gramas).

- Ventajas:
  - Simple de implementar
  - Reduce el sobreajuste de conteos altos en n-gramas raros
- Desventajas:
  - El valor del descuento δ necesita ajustarse correctamente para cada contexto
  - Puede requerir un modelo de respaldo efectivo para asignar probabilidades a eventos no observados.

2. Kneser-Ney Smoothing

Kneser-Ney Smoothing es una técnica de suavizado avanzada, considerada como una de las métodos más efectivas para modelar n-gramas. No solo toma en cuenta los conteos absolutos, sino también la distribución de los contextos en los que aparece un n-grama. Se basa en dos ideas fundamentales:

1. Suavizado basado en el número de contextos: Se le da más peso a las palabras que aparecen en muchos contextos diferentes, no solo en palabras que tienen un alto conteo.

2. Probabilidades de respaldo jerárquicas: Se utilizan probabilidades de respaldo calculadas en función de las frecuencias de los n-gramas del menor orden

- Definición:
  La probabilidad de un n-grama suavizado con Kneser-Ney se define como:

$$P(w_n|w_{1,\dots,}w_{n-1}) = \frac{\max(c(w_{i,\dots,}w_n)-d,0)}{c(w_{1,\dots,}w_{n-1})} + \lambda(w_{1,\dots,}w_{n-1}) \rightarrow$$

$$\rightarrow \cdot P_{backoff}[w_n|w_{2,\dots,}w_{n-1}).$$

donde d es un parámetro de descuento (puede ser constante o adaptativa), y la diferencia principal respecto al suavizado absoluto se encuentra en el cálculo de $P_{backoff}(w_n|w_{2,\dots,}w_{n-1})$

$$P_{backoff}(w_n|w_{2,\dots,}w_{n-1}) = \frac{|\{w_{n-1} : c(w_{n-1},w_n)>0\}|}{\sum_w |\{w_{n-1} : c(w_{n-1},w)>0\}|},$$

que representa la probabilidad de respaldo basada en el número de contextos distintos donde aparece la palabra w_n

- Ventajas:
  - Captura la diversidad contextual, lo que mejora significativamente el rendimiento
  - Es particularmente efectivo para modelos de lenguaje en los que las palabras con baja frecuencia son comunes (por ejemplo, en corpus pequeños o en problemas de escasez de datos).

- Desventajas:
  - Es más complejo de implementar que otros métodos de suavizado.
  - Requiere más cálculos, lo que puede ser computacionalmente costoso para corpus grandes.

Comparación

Ambas técnicas intentan resolver el problema de la probabilidad cero para n-gramas no observados, pero de distintas maneras:

- Absolute Discounting: asigna una parte de la probabilidad total a los n-gramas no observados restando un valor fijo de las cuentas observadas

- Kneser-Ney Smoothing: es un método basado en el contexto y el respaldo jerárquico, que toma en cuenta tanto las frecuencias como la diversidad contextual de los n-gramas.

En general, Kneser-Ney suele ser preferido en aplicaciones prácticas debido a su efectividad en diversos corpus y contextos, especialmente para n-gramas de alta variabilidad.