Investigación N-Grams II

Problemática:
El desafío central en modelos de lenguaje es el manejo de eventos raros o no observados. Cuando ciertas secuencias no aparecen en el conjunto de entrenamiento, pero podrían aparecer en el futuro, provoca que las probabilidades sean incorrectamente a cero.

Absolute Discounting o Kneser-Ney Smoothing son métodos que buscan distribuir las probabilidades de una manera más realista.

Absolute Discounting:
Resuelve este problema reduciendo o descontando una cantidad fija de probabilidad de los eventos observados y redistribuyendo ese valor a los eventos no observados.

$$P_{abs}(w_i \mid w_{i-n+1} \cdots w_{i-1}) =$$

$$\frac{\max\{C(w_{i-n+1} \cdots w_i) - D, 0\}}{\sum C(w_{i-n+1} \cdots w_i)} +$$

$$(1 - \lambda_{w_{i-n+1} \cdots w_{i-1}}) P_{abs}(w_i \mid w_{i-n+2} \cdots w_{i-1})$$

D = parametro desc.

λ = es un factor de normalización

Kneser - Ney Smoothing:
Le da un mayor peso a las palabras que se repiten en muchos contextos, incluso si aparecen menos en el corpus.

Tiene la misma expresión matemática, sin embargo ahora P es sustituido por PKN que toma en cuenta la cantidad de apariciones en diferentes contextos.

Ejemplo:

Considerando "el gato come" en un corpus en español. Si nunca se ha observado esa secuencia, los métodos de suavizado reasignarán una parte de la probabilidad de las secuencias observadas como "el gato duerme" o "el perro come", para asignar una probabilidad diferente de cero a "el gato come". Kneser-Ney considera que la palabra "gato" y "come" aparecen en muchos contextos por lo que asignaría una probabilidad más ajustada.