Modelos de lenguaje

Jordy Joaquin Cuan Robledo [1], Eduardo Cantorán Flores [1],
Rafael Pérez Aguirre [1]
[1] Facultad de Ciencias de la Computación, Benemérita Universidad
Autónoma de Puebla, México

Modelos de lenguaje

Un modelo de lenguaje es una representación abstracta de un lenguaje. Una aproximación al lenguaje real.
Existen dos modelos estadísticos, predictivos y explicativo.

Una parte útil del conocimiento necesario para la predicción de palabras o letras puede ser obtenido a través de técnicas estadísticas. Calculando la probabilidad de una secuencia, la posibilidad de co-ocurrencia de dos letras o palabras.

Aplicaciones de los modelos de lenguaje.

Podemos aplicar los modelos de lenguaje para diversas aplicaciones como:

- Reconocimiento de voz
- Reconocimiento de escritura
- Corrección de ortografía
- Sistemas de traducción
- Reconocimiento óptico de caracteres

Aproximación de palabras de la lengua natural

Aproximación de orden cero: la secuencia de letras es independiente entre sí y todas son igualmente probables.

Aproximación de primer orden: Las letras son independientes, pero se producen de acuerdo a la frecuencia del texto dado.

Aproximación de segundo orden: La probabilidad de que aparezca una letra depende la probabilidad de la letra anterior.

Aproximación de tercer orden: a probabilidad de que aparezca una letra depende de la probabilidad de las dos letras anteriores.

Los lenguajes que provienen de la misma familia son más similares entre sí que aquellos que provienen de otros lenguajes. Su similitud está basada en sílabas. En la siguiente tabla podemos observar la similitud que existe entre las sílabas de lenguas romance.

Language	The percentage covered by the first ··· syllables						No. syllables	
	100	200	300	400	500	561	type	token
Latin	72%	86%	92%	95%	98%	100%	561	3922
Romanian	63%	74%	80%	84%	87%	90%	1243	6591
Italian	75%	85%	91%	94%	96%	97%	803	7937
Portuguese	69%	84%	91%	95%	97%	98%	693	6152
Spanish	73%	87%	93%	96%	98%	99%	672	7477
Catalan	62%	77%	84%	88%	92%	93%	967	5624
French	48%	61%	67%	72%	76%	78%	1738	5691

La regla de la cadena

Existen modelos que nos permite calcular la probabilidad de una oración.
Predicción de palabras: Sencillo vs Inteligente

Sencillo: cada palabra seguida de otra palabra sobre la probabilidad
Tomando en cuenta que V es el total de palabras dentro de un texto, tenemos que la probabilidad de que la oración S de tamaño n es igual a 1/V * 1/V * …. * 1/V

Más inteligente: probabilidad de cada palabra siguiente está relacionado con frecuencia de palabras (unigramas)
La posibilidad de la oración S (es igual a la probabilidad de cada palabra) = P(w1) * P(w2) * … P(wn)
Asume que la probabilidad de cada palabra es independiente de la probabilidad de las demás palabras

Aún más inteligente: mira la probabilidad dada por las palabras anteriores (N-Gramas)
La probabilidad de cada palabra (es igual a la probabilidad de todas la palabra por la probabilidad de todas las palabras anteriores) = S = P(w1) × P(w2|w1) × ... × P(wn|wn-1)
Asume la probabilidad de que cada palabra es dependiente de las otras palabras.

Markov Assumption

Con el modelo de Markov Assumption podemos calcular la probabilidad de algo sin tener que utilizar muchos elementos pasados, para utilizar BiGramas necesitamos apicar la siguiente formula:

$$P(w1n) \approx \Pi k=1\ n\ P(wk|wk-1);\ w0 = \text{<start>}$$

EL orden de un modelo Markov se basa en la longitud del contexto previo, un bigrama es de primer orden, un trigrama es de segundo orden y así sucesivamente.

Un modelo de N-Grama utiliza la palabra N-1 anterior para predecir la siguiente.

$$P(Wn \mid Wn\text{-}N+1\ Wn\text{-}N+2 \ldots Wn\text{-}1\ )$$

Unigramas: P(perro)
Bigramas: P(perro | gran)
Trigramas: P(perro | el gran)
Cuatrigramas: (perro | persigue el gran)

$$\text{N-gram: } P(wn|w1n\text{-}1n\ ) \approx P(wn|wn\text{-}N+1n\text{-}1)$$

$$\text{Bigram: } P(w1n) \approx Pw \mid w$$

Técnicas de suavizado para modelos de lenguaje

Existen veces que aunque el corpus sea lo suficientemente grande, aparecerán N-gramas nuevos. En estos casos necesitamos utilizar un suavizado para que el resultado no se vea afectado.

Suavizado Add-One
Se tiene que agregar uno a cada N-Grama

$$P(wn|wn\text{-}1) = C(wn\text{-}1wn)/C(wn\text{-}1)$$
$$P(wn|wn\text{-}1) = [C(wn\text{-}1wn) + 1] / [C(wn\text{-}1) + V]$$