UNIVERSIDAD PERUANA DE CIENCIAS APLICADAS

Inteligencia Artificial

Modelos de basados en Corpus

- Hugo David Calderón
- Willy Ugarte Rojas
- Jorge Valverde Rebaza

Modelos basado en Corpus

Es un paradigma de Procesamiento de Lenguaje Natural, que utiliza modelos estadísticos y teorías de la información, analiza los corpus de textos para su procesamiento. El enfoque estadístico contrasta con los enfoques tradicionales como la traducción automática basada en reglas.

Ventajas sobre otros paradigmas

Los modelos basado en corpus o en estadística cuyo procesamiento se puede realizar con menos recursos por ejemplo para traducciones automáticas, se debe tener previamente documentos traducidos bilingües entre el par de lenguas, luego aplicando los métodos estadísticos se tendría un traductor automático sin utilizar reglas gramaticales.

Ventajas sobre otros paradigmas

- Uso de menos recursos como análisis de reglas.
- Implementando en modelos o fórmulas estadísticas no requieren un desarrollo manual de reglas lingüísticas
- Menor costo
- Se generaliza a otros idiomas

Ventajas sobre otros paradigmas

Hay mucho lenguaje natural desarrollado en formato legible que tales como escritos oficiales, textos, libros y otros corpus que pueden ser usados para procesamiento de lenguaje natural para varios propósitos.

Ventajas sobre otros paradigmas

Generalmente, los sistemas de traducción automática basado en estadísticas no están adaptados a ningún par específico de idiomas, pero es posible implementar en todo par de idiomas con la condición de existencia previa de corpus lingüístico.

Modelado del Lenguaje con N-Gramas

En la mayoría de las tareas de procesamiento de lenguaje natural es necesario identificar las sentencias o secuencias de palabras que ocurren con mayor probabilidad dado un contexto.

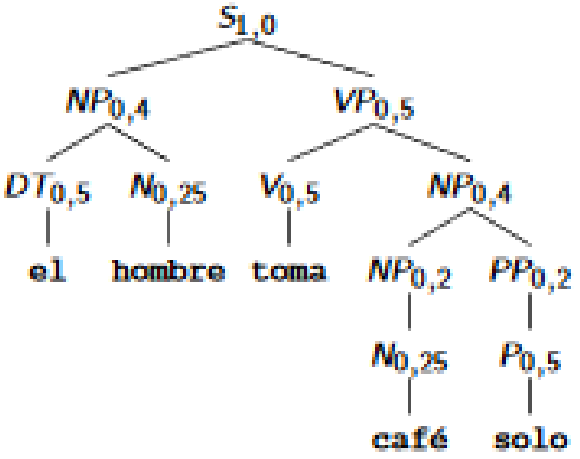
Modelo probabilístico basado en gramáticas

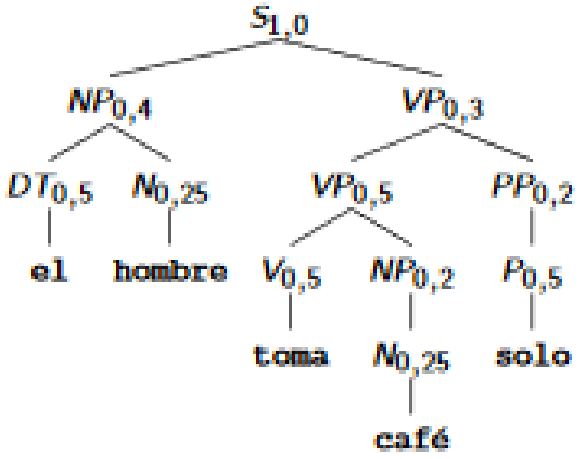
- Una gramática independiente de contexto probabilística es igual a una gramática independiente de contexto en la que cada regla tiene asociada una probabilidad.
- Estas gramáticas permiten calcular la probabilidad de una derivación sintáctica a partir de las probabilidades de todas las reglas que se han aplicado.
- La probabilidad de cada regla se aprende analizando colecciones de textos (corpus).
- De esta forma se intenta resolver la ambigüedad sintáctica: tómese el árbol de derivación mas probable.

Modelo probabilístico basado en gramáticas

S	$\Longrightarrow$	$NP\ VP$	1,0	
NP	$\Longrightarrow$	$DT\ N$	0,4	
		$\mid N$	0,2	
		$\mid NP\ PP$	0,4	
VP	$\Longrightarrow$	$V\ NP$	0,5	
		$\mid V$	0,2	
		$\mid VP\ PP$	0,3	
PP	$\Longrightarrow$	$P\ NP$	0,8	
		$\mid P$	0,2	
DT	$\Longrightarrow$	**el** | **los**	0,50	c.u.
N	$\Longrightarrow$	**hombre** | **amigos** | **café** | **leche**	0,25	c.u.
V	$\Longrightarrow$	**toma** | **toman**	0,50	c.u.
P	$\Longrightarrow$	**con** | **solo**	0,50	c.u.

Modelo probabilístico basado en gramáticas





- Probabilidad del primer análisis: 0,000025
- Probabilidad del segundo análisis: 0,0000187

Modelado del Lenguaje con N-Gramas

El objetivo de un modelo de lenguaje es calcular la probabilidad de ocurrencia de una sentencia que es originada por una secuencia de palabras que pasa a través de lo que se conoce como Noisy Channel.

Modelo de canal ruidoso

Modelo de canal ruidoso

Si dispusiéramos de una función que estimara la probabilidad de que se produzca una determinada entrada a partir de una salida, Pr(**e**|**f**), sería posible reconstruir la entrada más probable, **ê**, probando todas la posibles entradas de **e**, y seleccionando la de mayor puntuación.

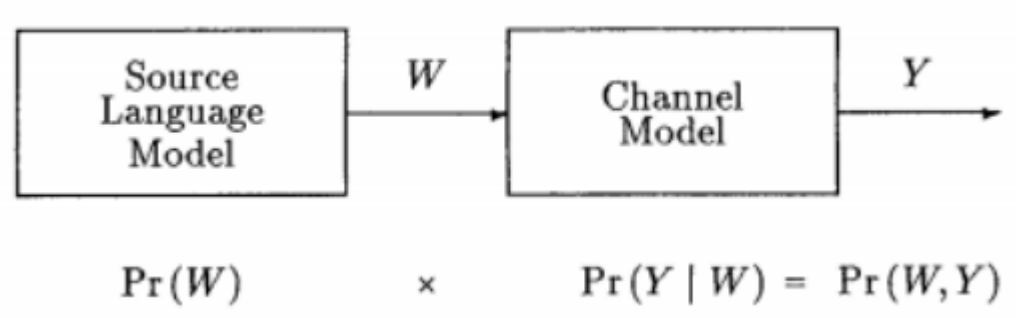
Modelo de canal ruidoso

El modelo de canal ruidoso se fundamenta sobre la función Pr(**e**|**f**). Aplicando el Teorema de Bayes.

$$\Pr(\mathbf{e} \mid \mathbf{f}) = \frac{\Pr(\mathbf{e}) \Pr(\mathbf{f} \mid \mathbf{e})}{\Pr(\mathbf{f})}$$

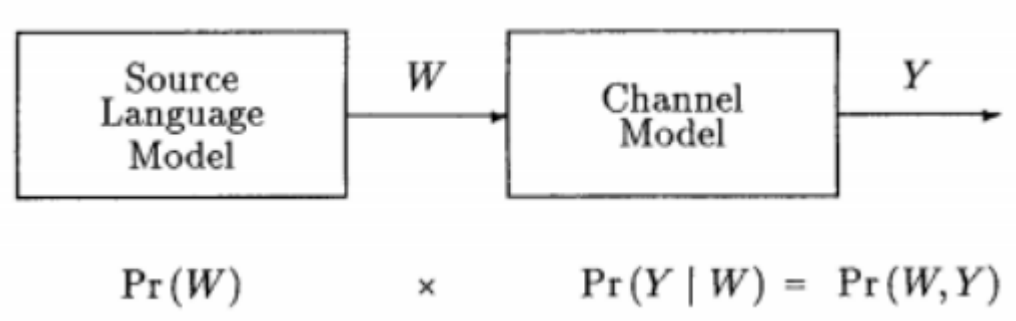
Modelo de canal ruidoso

El modelo es usado en aplicaciones y tareas como: reconocimiento de discurso, donde y es la señal acútica producida; en traducción automática de textos donde y es la sentencia en otro idioma; y en tareas de corrección automática de textos, donde y es la secuencia de caracteres emitida por un escritor imperfecto.



Modelo de canal ruidoso

En los tres casos anteriores, dada una secuencia de salida *y* se busca predecir la sentencia *w* que la origin´o (Peter E. Brown and Vincent J. Della Pietra, 1992).



Modelo de canal ruidoso

Los algoritmos que asignan probabilidades a una sentencia pueden ser también utilizados para calcular la probabilidad de la siguiente palabra dada una sentencia. Esto resulta de gran utilidad en tareas de part-of-speech-tagging (Daniel Jurafsky and James H. Martin, 2009).

Modelo de canal ruidoso

La probabilidad de una sentencia puede expresarse como

$$P(w_1 w_2 .. w_i) = \prod_i P(w_i | w_1 w_2 .. w_{i-1})$$
$$= P(w_1) P(w_2 | w_1) P(w_3 | w_1 w_2) .. P(w_i | w_1 w_2 .. w_{i-1})$$

Modelo de canal ruidoso

Ejemplo

$$
\begin{aligned}
P(\text{“}Lo\ esencial\ es\ invisible\ a\ los\ ojos''\text{)} \quad = \quad & P(lo) * P(esencial \mid lo) * \\
& P(es \mid lo\ esencial) * \\
& P(invisible \mid lo\ esencial\ es) * \\
& P(a \mid lo\ esencial\ es\ invisible) * \\
& P(los \mid lo\ esencial\ es\ invisible\ a) * \\
& P(ojos \mid lo\ esencial\ es\ invisible\ a\ los)
\end{aligned}
$$

Ven tajas e Inconvenientes

Ventajas

- Dan una idea probabilística de lo buena que es una derivación sintáctica de una frase, permitiendo decidir ante una ambigüedad
- Las reglas probabilísticas se pueden aprender a partir de un conjunto de ejemplos correctamente formado

Inconvenientes

- La probabilidad de una frase depende únicamente de la derivación sintáctica y no tiene en cuenta el contexto léxico: La frase el amigos toma hombre tiene la misma probabilidad que el hombre toma café.
- Las frases cortas tienen mayor probabilidad que las largas

Modelo de Ngramas

- Un modelo de n-gramas (modelo de lenguaje) intenta predecir la próxima palabra de una oración a partir de las N-1 anteriores.
- Objetivo: computar la probabilidad de que una palabra *w* ocurra luego de una secuencia previa *h*. P(*w* | *h*) Por ejemplo: P(conocimiento | como es de público) ¿Cómo podemos hacer?

Modelo de Ngramas

De manera general, las palabras $w_1 ... w_n$, su probabilidad se podría calcular

$$P(w_1 \ldots w_n) = P(w_1)P(w_2|w_1)\cdots P(w_n|w_1,\ldots,w_{n-1})$$

- Intuitivamente, cada $P(w_i|w_1,\ldots,w_{i-1},)$ es la probabilidad de que (en el lenguaje modelado) aparezca la palabra wi a continuacion de la secuencia $w_1,\ldots,w_{i-1}$
- Estas probabilidades se aprenden a partir de un corpus
- Pero en la practica es imposible saber la probabilidad de cada palabra condicionada a cada posible secuencia de palabras anteriores.
- Por esto, se toman determinadas suposiciones de independencia que simplifican el modelo (a costa de perder precisión)

Modelo de Ngramas

- Modelo *unigram*: Se asume independencia entre palabras consecutivas

$$P(w_1 \ldots w_n) = \prod_i P(w_i) \text{ con } P(w_i) = \frac{N(w_i)}{N}$$

Donde $N(w_i)$ es el número de ocurrencias de la palabra w_i en el corpus y N es el número total de palabras (incluyendo repeticiones)

- Modelo *bigram*: Se asume dependencia entre una palabra y la anterior, pero independencia con las demás

$$P(w_1 \ldots w_n) = P(w_1) \prod_i P(w_{i+1}|w_i) \text{ con } P(w_j|w_i) = \frac{N(w_i\ w_j)}{N(w_i)}$$

Donde $N(w_i\ w_j)$ es el número de ocurrencias de la secuencia (*bigram*) $w_i\ w_j$ en el corpus

  - Un *bigram* está formado por dos palabras consecutivas en el *corpus*

Modelo de Ngramas

- Modelo *trigram*: Se asume dependencia entre una palabra y las dos anteriores, pero independencia incondicional con las demás

$$P(w_1 \ \ldots \ w_n) = P(w_1)P(w_2|w_1)\prod_i P(w_{i+2}|w_{i+1}, w_i)$$

  - Un *trigram* está formado por tres palabras consecutivas en el *corpus*
- Modelo *n-gram*: Generalización de los modelos anteriores
  - Un *n-gram* está formado por *n* palabras consecutivas en el *corpus*
- En estos modelos probabilísticos, salvo el *unigram*, se tienen en cuenta relaciones contextuales léxicas, que no suelen aparecer en los modelos gramaticales

Corpus

Las probabilidades de un modelo de NGramas se obtienen a partir de un corpus de entrenamiento.

Idea: utilizar dos corpus, uno de entrenamiento y otro de prueba. Dado un problema: se recopila un conjunto de textos relevantes se divide en un corpus de entrenamiento (CE) y en un corpus de prueba (CP)

Fin