PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

FACULTAD DE LETRAS

Introducción a la lingüística computacional

César Antonio Aguilar

Facultad de Lenguas y Letras

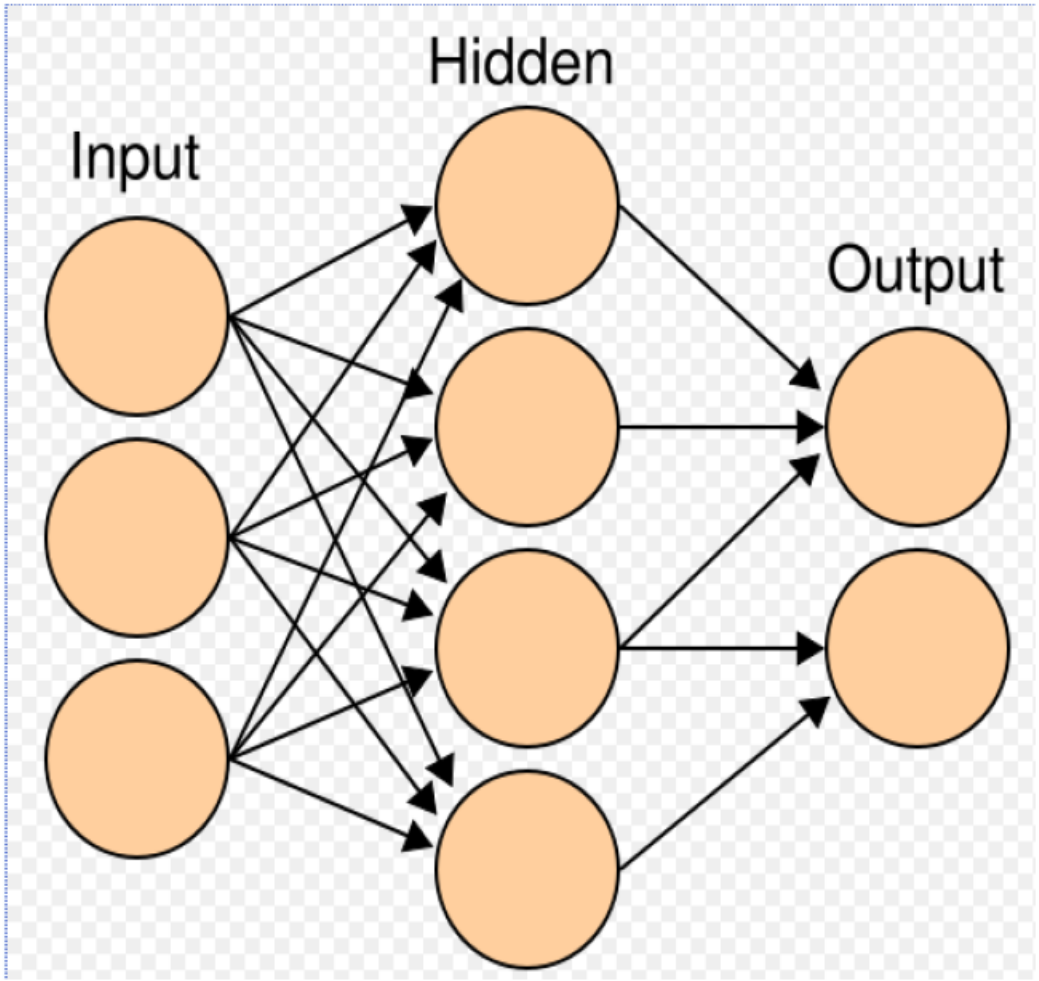
07/09/2017

Cesar.Aguilar72@gmail.com

Probabilidades + reglas lingüísticas

Conforme se ha avanzado en PLN, lo que parece que funciona hasta hoy son métodos híbridos que vinculen tanto conocimiento racional como empírico.

Empero, construir estos modelos no es fácil: dada la complejidad del lenguaje natural, se necesitan que éstos sean capaces de trabajar al unísono con ambos enfoques, y que sean muy finos a la hora de hacer el análisis.



Entropía (1)

Ya sabemos que una lengua natural es compleja. Sin embargo, ¿el hecho de que sea compleja la convierte en un sistema caótico? ¿Por qué deducimos que no lo es?

Hipótesis:
supongamos que cualquier lengua natural, si no la conocemos, nos puede parecer un fenómeno azaroso de entrada.
¿Podemos calcular cuán azarosa o no es su comportamiento?

Experimento:
pensemos en los átomos que hay en una botella de oxígeno. Si quisiéramos sacar una "foto" de cómo se configuran estos átomos, veríamos que su comportamiento, al ser un gas, es bastante desordenado.

Entropía (2)

Si viéramos cada una de estas fotos, éstas nos describen un comportamiento particular, de modo que la Foto 1 no se parece mucho a la Foto 2, y la Foto 3 tiene menos parecido con las Fotos 1, 2, etc.

Cada foto es un *microestado* que describe un comportamiento particular de nuestros átomos cada vez que se configuran dentro de nuestra botella.

Ahora, vamos a hacer una pequeña modificación: supongamos que nuestras fotos las sacamos a una temperatura de 25º C. ¿Tendrá el mismo comportamiento a unos -220º C?

Entropía (3)

Cuando hacemos esto, lo que estamos haciendo es analizar nuestro fenómeno dentro de un *macroestado*, es decir, dentro de un contexto que estamos regulando para nuestras evaluaciones (p. e., estamos controlando la temperatura de la botella).

La idea es que nuestro *macroestado* está asociando varios *microestados* de los átomos de óxigeno.

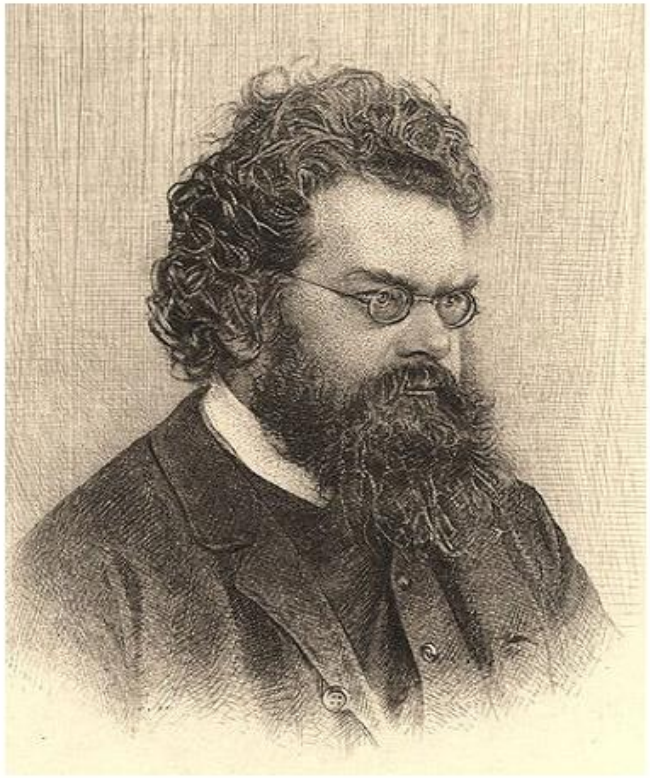
Lo interesante es que si cambiamos nuevamente la temperatura, p. e., la elevamos a unos 80º C, entonces nuestros átomos vuelven a enloquecer.

Entropía (4)

Posible explicación: parece que hay una relación directa entre cuá fría o caliente esté la botella con nuestros átomos para poder determinar qué tan azarosa o no es su organización.

Si quisieramos hacer predicciones, diríamos algo como: es más fácil que deduzcas una estrutura atómica del oxígeno cuando está congelado, que cuando está ardiendo.

En palabras más elegantes, los distintos microestados de los átomos del óxigeno tienden a agruparse en un solo macroestado cuando se congelan; en tanto que ocurre lo contrario si aumentamos la temperatura.



Ludwig Boltzmann
(1844-1906)

Entropía (5)

Pregunta: Si tengo un determinante, ¿con qué es más probable que se ligue?

(A) Con un nombre

(B) Con un verbo

(C) Con un adjetivo

(D) Con cualquiera de los tres previos

(E) NPI

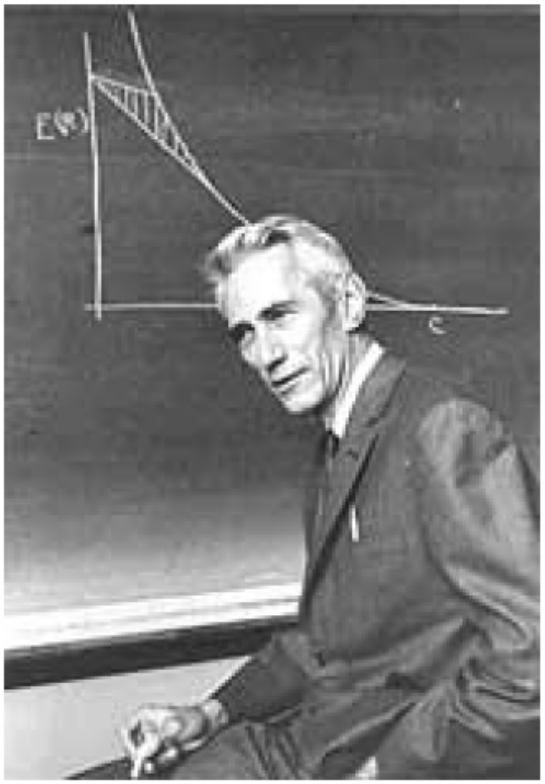
	word	count	frequency
1	the	69903	0.068271
2	of	36341	0.035493
3	and	28772	0.028100
4	to	26113	0.025503
5	a	23309	0.022765
6	in	21304	0.020807
7	that	10780	0.010528
8	is	10100	0.009864
9	was	9814	0.009585
10	he	9799	0.009570
11	for	9472	0.009251
12	it	9082	0.008870
13	with	7277	0.007107
14	as	7244	0.007075
15	his	6992	0.006829
16	on	6732	0.006575
17	be	6368	0.006219
18	s	5958	0.005819
19	I	5909	0.005771
20	at	5368	0.005243

Teoría de la información (1)

Antecendetes: Para responder a esta pregunta, podemos retomar la *teoría de la información* formulada por Shannon.

La teoría información investiga la probabilidad de sucesos inciertos, tratando de cuantificar numéricamente cuanta información aporta cada pista o hecho conocido que ayude a reducir la incertidumbre.

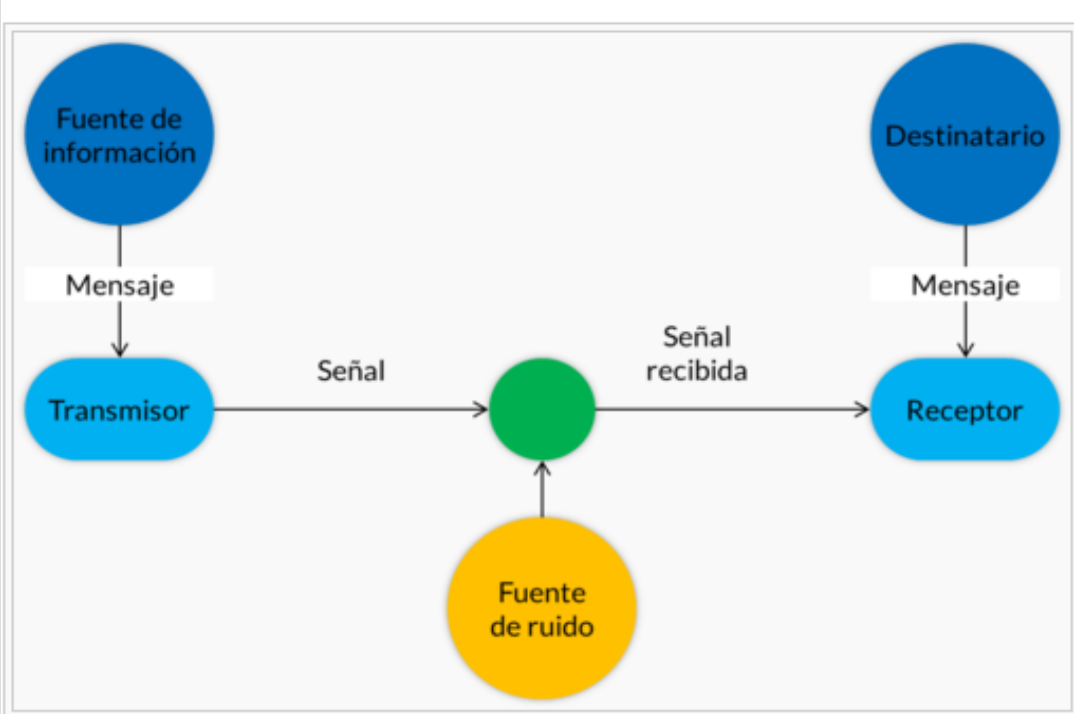
Por eso la información encerrada en un cierto "pedazo de conocimiento" es una función de las probabilidades de cada suceso posible en un evento incierto.



Claude Shannon
(1916-2001)

Teoría de la información (2)

Esta teoría es un sistema general de la comunicación que parte de una fuente desde la cual, a través de un transmisor, se emite una señal que viaja por un canal, pero a lo largo de su viaje puede ser interferida por algún ruido. La señal llega a un receptor que decodifica la información convirtiéndola posteriormente en mensaje que pasa a un destinatario. La teoría trata de determinar la forma más económica, rápida y segura de codificar un mensaje, sin que la presencia de algún ruido complique su transmisión.



Teoría de la información (3)

Matemáticamente, podemos hacer una fórmula que represente la idea anterior, esto es:.

$$I = I(p_1, \ldots, p_n | \text{conocimiento})$$

Donde las p_i son las probabilidades de ocurrencia de cada uno de los sucesos posibles compatibles con el conocimiento cierto que tenemos.

La teoría de la información no puede decirnos si cierto conocimiento es verdadero o falso, sólo cuantificar numéricamente cuanto es ese conocimiento en relación a la incertidumbre existente bajo la suposición de que el conocimiento que tenemos es verdadero.

Teoría de la información (4)

Retomando el problema sobre cómo puedo predecir cuál palabra va a seleccionar un determinante, si deseo verlo como un fenómeno con cierto grado de entropia, tenemos dos escenarios:

Alto nivel de entropia:

Los determinantes seleccionan cualquier palabra, por tanto no podemos hacer ninguna predicción segura sobre qué comportamiento asumen dentro de una cadena de palabras.

Bajo nivel de entropia:

Los determinantes seleccionan algunas palabras sobre otras, por tanto podemos predecir, asumiendo cierto margen de error, cuál será su comportamiento dentro de una cadena de palabras.

Teoría de la información (5)

Ahora, pensando en lo que vimos sobre probabilidades condicionales, podemos expresar lo siguiente:

Dada una variable aleatoria X que toma valores *x1, x2, ..., xn* en un dominio de acuerdo con una distribución de probabilidad, podemos definir el valor esperado de X como la suma de los valores ponderados con su probabilidad, esto es:

$$E(X) = p(x1)X(x1) + p(x2)X(x2) + \ldots p(xn)X(xn)$$

Teoría de la información (6)

En esta propuesta, partimos de la idea de que el azar existe en el mundo, pero entre más reduzcamos ese azar, podemos descubrir que hay relaciones estrechas entre eventos en aparencia independientes. Formalmente, pensamos en términos de:

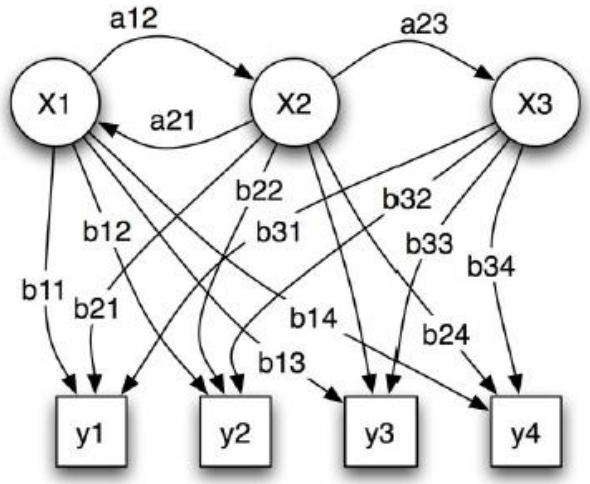
Significancia: cuanto más improbable es un evento más información lleva, es decir: $P(x1) > P(x2) \Longrightarrow I(x2) < I(x1)$

Aditividad: si x1 y x2 son eventos independentientes, esto es: $I(x1x2) = I(x1) + I(x2)$

Máquinas de Markov (1)

Una vez que vimos qué es la entropia como un método de análisis probabilístico, consideremos ahora la posibilidad de diseñar un autómata (con sus estados y transiciones), el cual es capaz de hacer inferencias a partir de métodos probabilísticos.

A esta máquina la conocemos por el nombre de *máquina de Markov.*



Andrei Markov (1856-1922)

Máquinas de Markov (2)

Siendo rigurosos, podemos definir una máquina (o cadena) de Markov en los siguientes términos:

Es un proceso estocástico discreto que cumple con la propiedad de Márkov, es decir, si se conoce la historia del sistema hasta su instante actual, su estado presente resume toda la información relevante para describir en probabilidad su estado futuro.

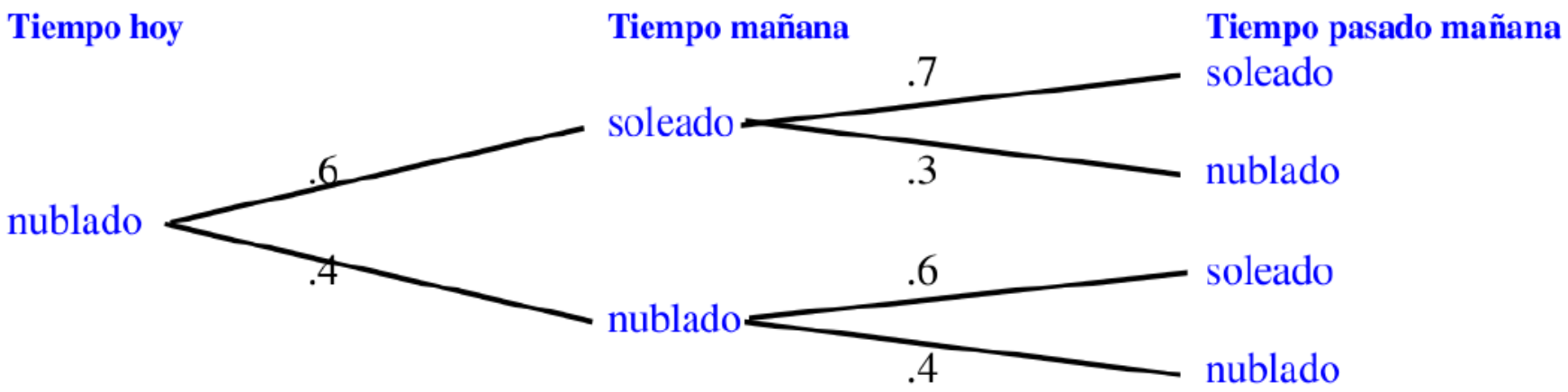
Máquinas de Markov (3)

Si nos ponemos más líricos, podemos decir que se trata de una máquina capaz de recordar el pasado asociado a un evento dado, y con ello puede de mirar y predecir el futuro. O sea:

$$P\left(X_{n+1}=x_{n+1}\middle|X_n=x_n,X_{n-1}=x_{n-1},\ldots,X_2=x_2,X_1=x_1\right)=P\left(X_{n+1}=x_{n+1}\middle|X_n=x_n\right).$$

Máquinas de Markov (4)

¿Por qué es conviene que las máquinas de Markov puedan reconocer el pasado? Porque esto las ayuda a tomar deciones, dada una probabilidad condicional: si ocurre un evento A, y este evento A influye en el evento B, entonces esmuy probable que ocurra B. Por ejemplo, ¿cómo predice el clima una de estas máquinas?



Máquinas de Markov (5)

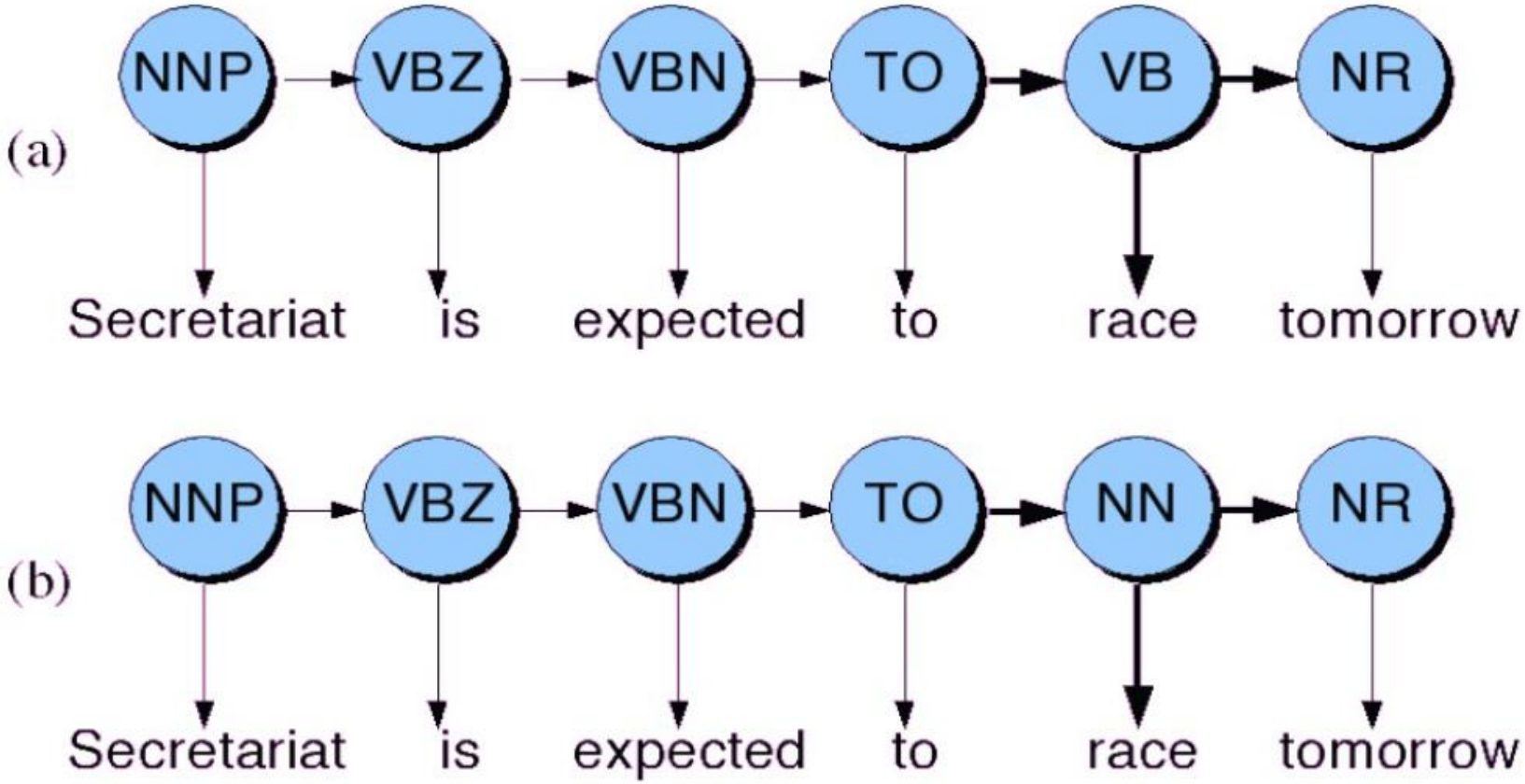
¿*Race* es un verbo, o es un nombre?

Secretariat/NNP is/VBZ expected/VBN to/TO **race**/VB tomorrow/NR

People/NNS continue/VB to/TO inquire/VB the/DT reason/NN for/IN the/DT **race**/NN for/IN outer/JJ space/NN

Máquinas de Markov (6)

Determinar qué etiqueta gramatical le atribuimos a *race* dentro de un corpus se conoce como *desambiguación*.



Máquinas de Markov (7)

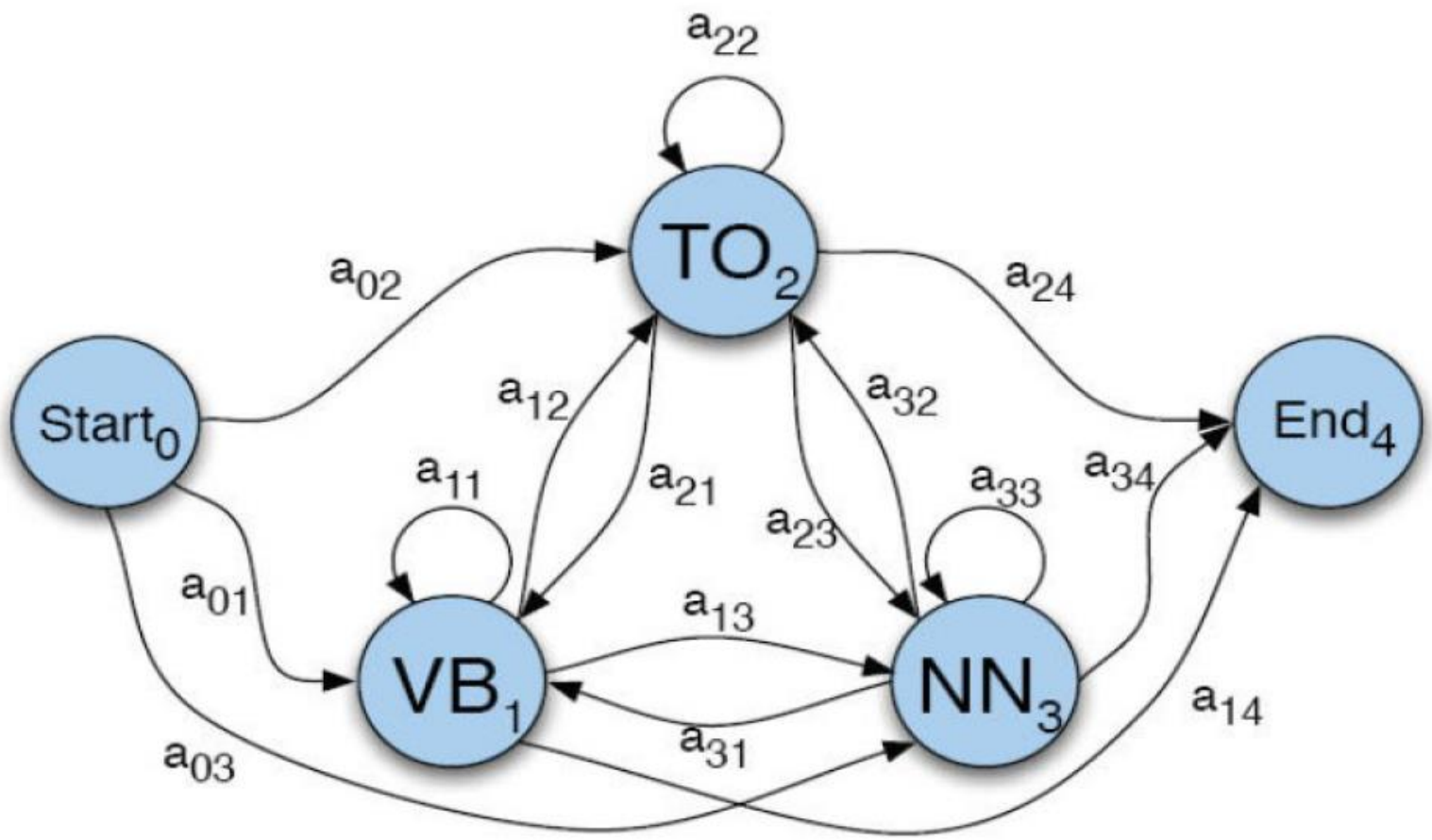
¡Y el ganador es...!

P(NN|TO) = .00047
P(VB|TO) = .83
P(race|NN) = .00057
P(race|VB) = .00012
P(NR|VB) = .0027
P(NR|NN) = .0012

P(VB|TO)P(NR|VB)P(race|VB) = .00000027
P(NN|TO)P(NR|NN)P(race|NN)=.00000000032

Máquinas de Markov (8)

Nuestra máquina de Markov sería la siguiente:



Máquinas de Markov (9)

Y estas son nuestras tablas de probabilidades:

	VB	TO	NN	PPSS
<s>	.019	.0043	.041	.067
VB	.0038	.035	.047	.0070
TO	.83	0	.00047	0
NN	.0040	.016	.087	.0045
PPSS	.23	.00079	.0012	.00014

	I	want	to	race
VB	0	.0093	0	.00012
TO	0	0	.99	0
NN	0	.000054	0	.00057
PPSS	.37	0	0	0

Gracias por su atención

Blog del curso:
http://cesaraguilar.weebly.com/introduccioacuten-a-la-linguumliacutestica-computacional.html