UNIVERSIDAD AUTÓNOMA DE QUERÉTARO

Seminario de análisis del discurso

Dr. César Antonio Aguilar
Facultad de Lenguas y Letras
20/09/2010

CAguilar@iingen.unam.mx

Estadísticas en corpus (1)

```
>>> reference = 'This is the reference data.  Testing 123.  aoacococ'
>>> test =      'Thos iz_the rifirenci data.  Testeng 123.  aoaeoeoe'
>>> print ConfusionMatrix(reference, test)
  |   . 1 2 3 T _ a c d e f g h i n o r s t z |
--+-----------------------------------------+
  |<8>. . . . 1 . . . . . . . . . . . . . . |
. | .<2>. . . . . . . . . . . . . . . . . . |
1 | . .<1>. . . . . . . . . . . . . . . . . |
2 | . . .<1>. . . . . . . . . . . . . . . . |
3 | . . . .<1>. . . . . . . . . . . . . . . |
T | . . . . .<2>. . . . . . . . . . . . . . |
_ | . . . . . .<.>. . . . . . . . . . . . . |
a | . . . . . .<4>. . . . . . . . . . . . . |
c | . . . . . .<1>. . . . . . . . . . . . . |
d | . . . . . .<1>. . . . . . . . . . . . . |
e | . . . . . .<6>. . . 3 . . . . . . . . . |
f | . . . . . .<1>. . . . . . . . . . . . . |
g | . . . . . .<1>. . . . . . . . . . . . . |
h | . . . . . .<2>. . . . . . . . . . . . . |
i | . . . . . 1 . . .<1>. 1 . . . . . . . . |
n | . . . . . .<2>. . . . . . . . . . . . . |
o | . . . . . .<3>. . . . . . . . . . . . . |
r | . . . . . .<2>. . . . . . . . . . . . . |
s | . . . . . .<2>. 1 . . . . . . . . . . . |
t | . . . . . .<3>. . . . . . . . . . . . . |
z | . . . . . .<.>. . . . . . . . . . . . . |
--+-----------------------------------------+
(row = reference; col = test)
```

Como comentamos al inicio de esta unidad, uno de los rasgos interesantes del análisis de corpus es poder contar con un soporte empírico importante para validar alguna hipótesis sobre un fenómeno lingüístico.

Estadísticas en corpus (2)

Con el ejercicio que realizaron dentro del corpus técnico del IULA, habrán visto lo pertinente que resulta usar algún método estadístico para calcular cuál es el grado de confianza que podemos esperar de un conjunto de datos para validar o no una hipótesis.

En terminología lo anterior es la regla: ¿cómo saber que una construcción sintáctica es un término? Simplemente calculando su frecuencia de uso dentro de documentos pertenecientes a un área técnica o científica.

Estadísticas en corpus (3)

En estos casos, el concepto de **patrón** alude precisamente a una estructura lingüística que, debido a su regularidad estadística, es útil para encontrar información relevante para resolver una consulta.

	items	noise	Examples: acceptable term candidates	Examples: noise
Individual queries:				
P + N	2990	13,5%	*Hinterachse, Vollgas*	*Bedeutung, Gegensatz*
P + ADJ	1755	46%	*gefiltert, verstellbar*	*unabhängig, vorhanden*
P + V	1603	31%	*beschleunigen, einspritzen*	*beibehalten, entscheiden*
S + N	6349	8%	*Geschwindigkeit, Produktion*	*Kenntnis, Wirklichkeit*
S + ADJ	2674	29,5%	*bleifrei, lieferbar*	*gemeinsam, unmittelbar*
DC + N	5376	2%	*Dachluke, Motorbremse*	*Nachteil, Umgang*
DC + ADJ	786	22%	*thermodynamisch, betriebswarm*	*nachhaltig, nutzlos*
DC + V	585	29%	*einbauen, herausfiltern*	*verkraften, aufteilen*
Abbr.	783	1%	*km/h/s, FCKW-frei, TE-24*	*MANN, BOSCH-Einspritzpumpe*
Combined Queries:				
P + S + N	2270	13%	*Sonderausstattung, Verstellhebel*	*Verhältnis, Beschreibung*
P + S + ADJ	1508	51%	*einstellbar, hochporös*	*entsprechend, abhängig*
P + DC + N	1713	6,5%	*Leergas, Innenraumluft*	*Aufwand, Vorteil*
P + DC + ADJ	371	28%	*hochbelastet, ungefiltert*	*vorhanden, entsprechend*
P + DC + V	469	28%	*zuschalten, ansaugen*	*bewirken, erhalten*
S + DC + N	4233	2,5%	*Lasersensor, Kraftübertragung*	*Erteilung, Schimmelbildung*
S + DC + ADJ	747	20,5%	*elektronisch, nachgeschaltet*	*typisch, wirkungsvoll*
P + S + DC + N	1318	4%	*Abgasleitung, Feststellbremse*	*Erfahrung, Aufteilung*
P + S + DC + ADJ	363	24%	*vollelektronisch, hochbelastbar*	*ermittelt, vorausgegangen*

Probabilidades (1)

Cuando manejamos enormes cantidades de datos lingüísticos (de cualquier tipo), si bien es cierto que muchas cosas las podemos deducir por sentido común, hay otras en donde requerimos determinar cuán probable o no es que tales datos reflejen un comportamiento determinado.

Un experimento sencillo: ¿cómo podemos calcular, en una partida de dados, cuál es el número más probable que salga en una tirada?

Delimitemos:

1. Un dado tiene 6 caras.
2. Cada cara representa un número del 1 al 6.
3. En teoría, en 6 tiradas me puede salir al menos un número del 1 al 6.

Probabilidades (2)

Para hacer nuestros cálculos, necesitamos una tabla como la siguiente:

Tiradas	Caras del dado					
1	1/6	1/6	1/6	1/6	1/6	1/6
2						
3						
4						
5						
6	1	2	3	4	5	6

De primera instancia, podemos plantear dos cosas:

1. Si el dado no tiene ningún truco, en teoría cada tirada (en una serie de 6) nos muestra un número concreto del dado.

2. Por el contrario, si el dado tiene algún truco, es muy posible que tienda a mostrar un número una buena cantidad de veces.

Probabilidades (3)

Dependiendo de una u otra opción, tenemos dos clases de probabilidades:

Probabilidad independiente o relativa: se da cuando planteamos la ocurrencia de dos eventos (a y B) que no mantienen ninguna relación entre sí.

$$P(A \text{ and } B) = P(A) \times P(B)$$

Nota: esto aplica para nuestro caso 1: cada vez que lancemos un dado, nos va a dar siempre una cara diferente.

Probabilidades (4)

Probabilidad condicional: se da cuando planteamos la ocurrencia de dos eventos (A y B) o más, tomando en cuenta que el evento A determina el evento B.

Para decirlo en otras palabras: la probabilidad de que ocurra el evento B dado el evento A se puede representar con la fórmula:

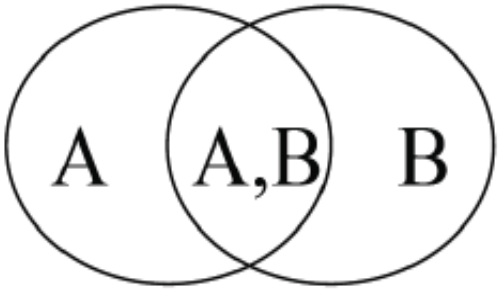
$$p(B|A) = p(A \cap B) / p(A)$$

Nota: esto aplica para nuestro caso 2, y justo son la clase de probabilidades que nos interesan.

Probabilidades (5)

Si nos ponemos más formales, una probabilidad condicional equivale a decir que:

$$P(A \mid B) = \frac{P(A,B)}{P(B)}$$



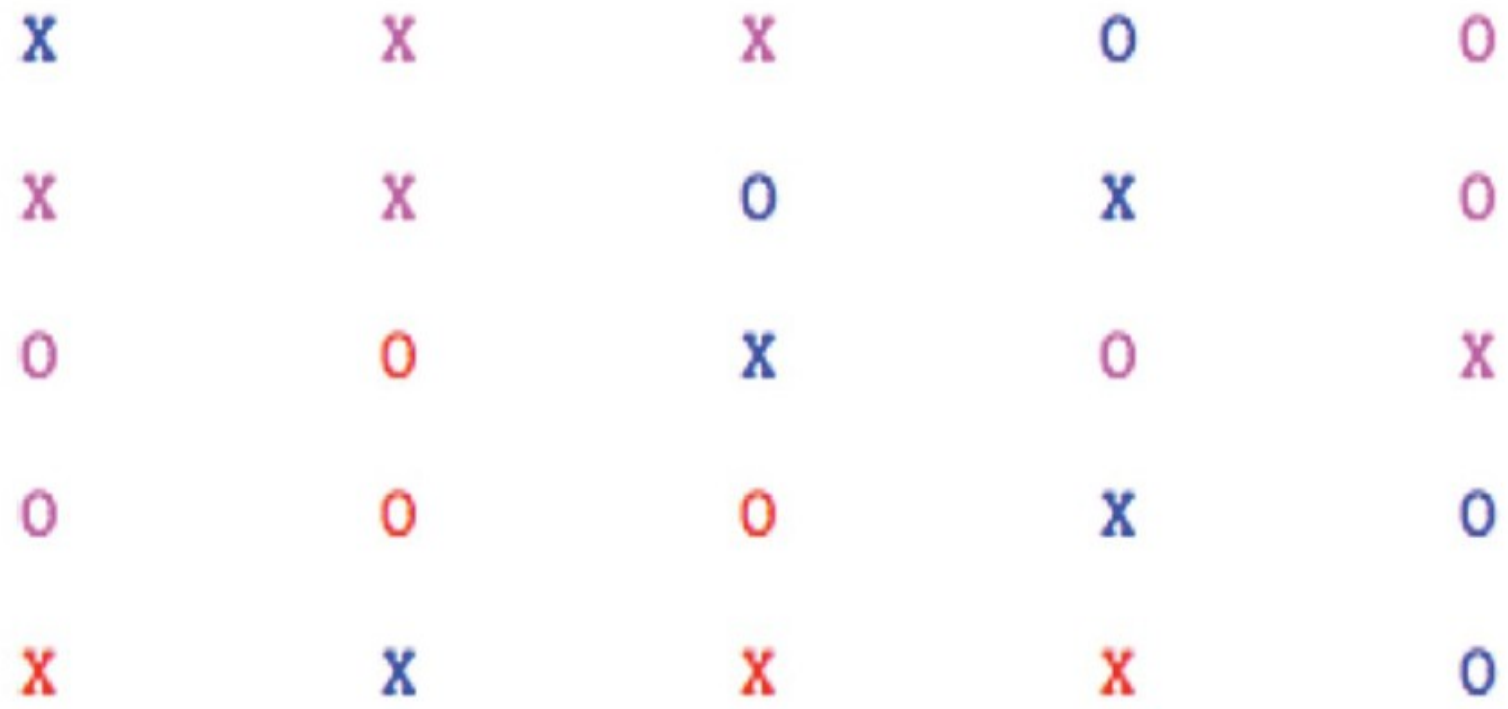
Recordando las matemáticas de la secundaria, esto se puede representar con un diagrama de Venn:

Hay que notar que: $P(A,B)=P(A|B) \cdot P(B)$

También: $P(A,B) = P(B,A)$

Probabilidades (6)

Veamos un ejemplo: con base en la probabilidad de que una figura sea una X, esto es, P(X), ¿cuántas probabilidades hay de que dada una X sea roja, es decir, P(X|roja)?



Probabilidades y corpus (1)

Cuando vemos el lenguaje desde un punto de vista probabilístico, podemos considerar que los fenómenos que ocurren en él son de carácter condicional, más que fruto del azar.

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Volviendo a algo que vimos la clase pasada: la detección de **n-gramas** es más que nada un proceso estadístico que ayuda al trabajo de inserción de etiquetas.

Probabilidades y corpus (2)

Las probabilidades son empleadas muchas tareas, p. e., reconocer relaciones sintácticas y semánticas entre palabras:

En este caso, si establecemos que Shakespeare es un autor, es porque tal atributo aparece asociado a dicho nombre con una frecuencia alta, de acuerdo con la información que nos brinda, p. e., el **British National Corpus**.

"...works by such **authors** as Herrick, Goldsmith, and **Shakespeare**."

"If you consider **authors** like **Shakespeare**..."

"Some **authors** (including **Shakespeare**)..."

"**Shakespeare** was the **author** of several..."

"**Shakespeare**, **author** of *The Tempest*..."

Shakespeare IS-A ***author*** (0.87)

Experimento (1)

Veamos cómo podemos aprovechar esta clase de métodos para hacer búsquedas en corpus. En el siguiente sitio electrónico:

http://labs.translated.net/

En esta página encuentran un extractor de términos en línea, desarrollado por Translated Labs, que es una empresa dedicada a dar servicios de ingeniería lingüística.

Con este extractor, lo que queremos es obtener posibles candidatos a términos dentro de un documento, en este caso, un artículo escrito en inglés.

Experimento (2)

El extractor nos muestra las 20 unidades que mostraron una alta frecuencia de uso que, por ello, pueden ser posibles candidatos a términos:

#	Extracted term	Score
1	verb	62%
2	definitional patterns	61%
3	definitional contexts	60%
4	defining verb	60%
5	como	59%
6	cociente entre el área	59%
7	con el intercambio entre	59%
8	buque y el área	59%
9	corpora	58%
10	onomasiological dictionary	57%
11	verbal predication	57%
12	adverb como	56%
13	con este trabajo	55%
14	differentia	55%
15	specialised corpora	55%
16	definition aquel recurso	55%
17	verbs	55%
18	definir	54%
19	defining verb lemmas	54%
20	person plural pronoun	53%

Podrán notar que no todos los posibles candidatos parecen ser buenos términos, pero algunos otros sí son buenas opciones.

Pregunta: en esta clase de tareas, ¿cómo podrían aprovechar probabilidades condicionales para explorar un corpus y obtener datos relevantes?

Experimento (3)

Veamos otro caso, y tratemos de resolver estas preguntas:

1. En un artículo escrito por Yorick Wilks, ¿a qué se refieren los términos “semantic primitive derivation”, “statistical nlp” y “modern statistical nlp” ?

2. Los dos últimos términos, ¿se refieren al mismo concepto, o aluden a dos conceptos claramente diferenciados?

3. ¿Con qué palabras se relaciona el término “corpora”?

4. Las palabras “syntactic” y “computational”, ¿son buenos ejemplos de términos?

Experimento (4)

Para tener una mejor idea, veamos los resultados:

#	Extracted term	Score
1	syntactic	66%
2	linguistics	63%
3	computational	62%
4	corpora	58%
5	wilks	58%
6	sparck jones	56%
7	syntactic structure	56%
8	semantic primitive derivation	55%
9	statistically-based empirical emphasis	54%
10	programming language prolog	54%
11	cat-dog sentence	54%
12	syntactic analysers	53%
13	dictionaries	53%
14	computational linguistics	53%
15	statistical nlp	53%
16	grammar parsing	53%
17	modern statistical nlp	53%
18	overtax syntactic	53%
19	computational semantics	52%
20	roget 's thesaurus	52%

Tarea

Hagan una extracción de términos usando un artículo de Mira Ariel (el archivo se llama *Ariel01.txt*, y lo pueden descargar desde la página del blog. Una vez que obtengan sus términos, traten de responden a las siguientes preguntas:

1. Los terminos que enlista el extractor, ¿son todos los que hay en el artículo de Ariel? Revisen el artículo y propongan otros que no aparezcan.

2. ¿Cuál es el grado de confianza que le darían al extractor? Esto es: ¿los términos que no reconoció, son muchos, son pocos? ¿Cómo harían esta evaluación?

3. El término *pronoum* parece que es muy productivo en el texto de Ariel, y de hecho permite construir otros con una estructura más compleja (esto es, se constituyen a partir de 2 o más palabras). ¿Qué otros términos podrían derivar usando *pronoum*? Propongan por lo menos 10 posibles nuevos términos.

Gracias por su atención

Blog del curso: http://discurso-uaq.weebly.com/