PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

FACULTAD DE LETRAS

Curso de traducción automática de lenguas naturales

César Antonio Aguilar

Facultad de Lenguas y Letras

06/11/2015

Cesar.Aguilar72@gmail.com

Creando un modelo del lenguaje (1)

Vamos a continuar en esta sesión con nuestra aproximación al uso de **Sketch Engine** para hacer un análisis con corpus paralelos.
Lo primero que podemos intentar hacer es generar un listado de palabras con sus frecuencias de aparición:

Word list

Corpus: EUROPARL7, en

Page 1 Go Next >

word	Freq
the	3,881,912
of	1,973,758
to	1,823,837
and	1,552,787
in	1,192,134
that	947,926
is	912,855
a	902,995
for	615,116
I	598,599
on	544,255
this	487,606
be	487,350
we	468,099
are	404,742
have	386,611
it	382,623
not	366,035
as	349,809
with	344,547
which	336,025
European	320,823
The	296,021

Creando un modelo del lenguaje (2)

Sketch Engine nos permite elaborar diferentes listad de palabras, tomando en cuenta atributos como reconocerlas por su forma (*tokens*), su lema o su etiqueta morfosintáctica. Veamos:

Word list

Corpus: **EUROPARL7, en**

Page 1 Go Next >

lemma	Freq
the	4,177,530
be	2,263,879
of	1,980,692
to	1,823,837
and	1,565,324
in	1,292,558
that	994,449
a	922,617
have	698,205
for	636,970
we	628,700
I	598,604
this	590,119
on	563,524
it	514,448
as	390,962
not	369,044
with	355,348
which	336,525
European	323,308

¿Todas las palabras son relevantes? (1)

Hasta el momento, estamos viendo un texto como una colección de palabras útil, la cual nos ayuda a determinar cuál es su contenido temático de información. Sin embargo, no todas las palabras nos ayudan a inferir esta información.

George Kingsley Zipf (1902-1950)

Un lingüista norteamericano, G. K. Zipf, a partir de varios estudios de comparación entre lenguas, determinó que no necesariamente las palabras más recurrentes son las que tienen mayor contenido significativo.

Del mismo modo, las palabras que muestran un bajísimo índice de uso, tendrán a mostrar frecuencias de aparición muy bajas.

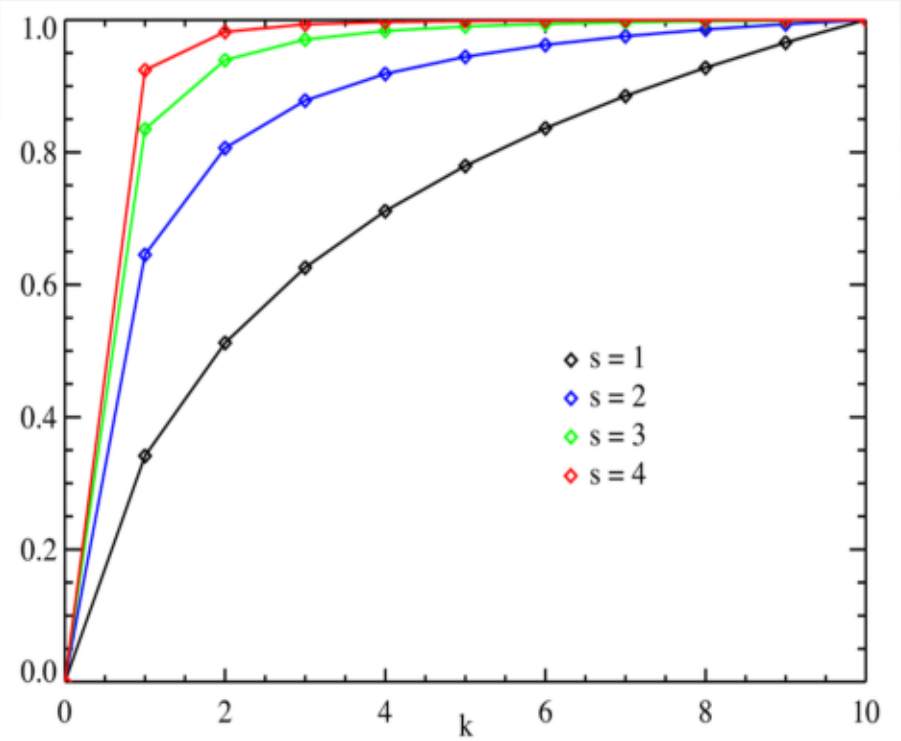
Ergo, lo que parece interesante es que, entre los dos extremos, podemos identificar palabras que nos aportan algún significado relevante.

¿Todas las palabras son relevantes? (2)

La llamada *Ley de Zipf* nos ayuda entonces a identificar qué posibles palabras realmente den cuenta del tema de un texto, y cuáles simplemente desempeñan meras funciones gramaticales.

La ley de Zipf tiene la ventaja de que, si trabajamos con grandes cantidades de palabras, de tal modo que nos brinda una "radiografía" confiable sobre el vocabulario relevante de un documento (o de varios documentos).

Lo interesante, como hemos planteado desde un principio, es que no necesitamos "leer" cada uno de nuestros documentos para saber de qué tratan.



¿Todas las palabras son relevantes? (3)

¿Cómo funciona la ley de Zipf? A grandes rasgos, lo que propone es que:

$$P_n \sim 1/n^a$$

Lo que expresa esta fórmula es lo siguiente: *Pn* representa la frecuencia de una palabra ordenada n-ésima, mientras que *a* representa un valor casi cercano a 1. Esto quiere decir que el segundo elemento se repetirá aproximadamente con una frecuencia de 1/2 de la del primero, y el tercer elemento con una frecuencia de 1/3 y así sucesivamente. Otra manera de expresarlo es:

$$f(k;s,N) = \frac{1/k^s}{\sum_{n=1}^{N}(1/n^s)}.$$

Aquí lo que representamos es: *K* es el número de elementos (o palabras), *N* es el rango de cada uno de estos elementos, y *s* es el valor que asume su exponente respecto a su distribución. La suma de todos estos elementos nos permite delimitar cuáles son los que tienen una mayor distribución.

Experimento (1)

Tratemos de hacer nuestra primera tarea de corpus paralelo, analizando un conjunto de 10 artículos de medicina, en español e inglés, usando **Sketch Engine**. Así, lo que hay que crear primero es nuestro modelo del lenguaje para ambos idiomas. Para ello, primero tenemos que acceder a esta colección. Lo que hay que hacer entonces es:

1. Entrar a la siguiente cuenta en el sitio de **Sketch Engine: CAGUILAR**

2. Seleccionar la opción *My Own*, y luego elegir el corpus denominado *Parallel_Corpus_Medicine*

Experimento (2)

Este corpus está conformado por los siguientes documentos:

Parallel_Corpus_Medicine

parallel_corpus_medicine

Add new file | Add data from web using WebBootCaT | Compile corpus | Search corpus

#	Original file	Plain text	Vertical	Tokens	Owner	
1	English_Medical_Corpus01.txt	✓	✓	33,846	Dr. César Antonio Aguilar	
2	Spanish_Medical_Corpus01.txt	✓	✓	44,224	Dr. César Antonio Aguilar	

Experimento (3)

Una vez hecho esto, la siguiente parte es hacer la **compilación** de su corpus. Para ello, nos apoyará la Dra. Olga Acosta. La idea entonces es:

I. Etiquetar su corpus de medicina, tanto e inglés como en español, usando anotado morfosintáctico.

II. Una vez hecho este etiquetado, hay que generar un listado de palabras para cada corpus.

III. Traten de generar distintos conteos de palabras, contrastando las opciones que les da Skecth Engine

IV. Finalmente, traten de determinar, considerando la Ley de Zipf, cuáles son las palabras representativas de sus corpus. ¿Hay coincidencias, o hay divergencias, a pesar de ser el mismo texto vertido en dos idiomas?

Gracias por su atención

Blog del curso:
http://cesaraguilar.weebly.com/traduccioacuten-automaacutetica.html