Primera aproximación a un extractor terminológico independiente de la lengua y del dominio temático.

Transcripción del Seminario IULATerm
dictado por:
Rogelio Nazar
en la
Universidad Pompeu Fabra
el 22 de Abril, 2008

rogelio.nazar@upf.edu

1. Introducción

La temática del seminario que voy a ofrecerles hoy, o por lo menos la segunda parte, tiene una íntima relación con algo que ha sido objeto de discusión recientemente en el grupo, que es el problema de la función denominativa de una expresión. La extracción automática de terminología especializada es, por demás, un tema de gran interés para el IULA. La particularidad de la propuesta que les traigo aquí es que se trata de un sistema independiente de la lengua y del dominio temático y eso plantea cierta novedad respecto a otras propuestas que habíamos estudiado hasta ahora (Vivaldi, 2001). Presentaré unas investigaciones que si bien están en un estado algo prematuro, pueden servir ya para dar una idea de cuál puede ser una metodología adecuada para trabajar con documentos de cualquier lengua y dominio, sin la necesidad de llevar a cabo un procesamiento lingüístico de los textos ni disponer de acceso a fuentes de conocimiento externas como ontologías o diccionarios.

Hay motivos por un lado de orden práctico para excluir estas fuentes de conocimiento externo, por la posible dificultad que podemos encontrar en conseguirlas o en adaptar nuestro algoritmo para que acceda a ellas. Pero también hay consideraciones de orden teórico, y es que hoy en día es posible que podamos prescindir de estas fuentes. En la actualidad, disponiendo de un corpus tan grande como Internet, se puede prescindir de conocimiento no sólo por apego a la simplicidad, sino porque incluso podríamos obtener mejores resultados -o al menos esa es la sospecha que quiero despertar- cuando planteamos la extracción terminológica como un problema puramente matemático.

Es oportuno el seminario también porque recientemente tuve acceso a los comentarios generados por los estudiantes de esta casa, como *feed-back* de uso del programa Terminus. Me sorprendió advertir que la mayoría había interpretado que Terminus era un programa de extracción de terminología, y esto es un error muy grande ya que no se puede pensar que el trabajo de ordenar cadenas de palabras o enegramas por frecuencia decreciente sea el trabajo de un extractor. Un extractor no se puede limitar a eso, esta sería una estrategia muy ingenua, condenada irremediablemente al fracaso. Entonces me parece percibir la necesidad de desarrollar un programa que sea simple y que tenga capacidad de adaptación.

Todavía hace falta algún tiempo para presentar un prototipo que pueda utilizarse, sin embargo, el diseño del algoritmo está bastante avanzado. Lo que falta es someterlo a una evaluación intensiva en distintos contextos lingüísticos y temáticos para tener datos fiables acerca de su desempeño. El resto de este seminario está organizado de la siguiente manera: en primer lugar presentaré algunas estrategias estadísticas más o menos simples que pueden utilizarse con el propósito de extraer o

resaltar la terminología especializada de un texto o un corpus. En esta primera parte hablaré rápidamente de algunas medidas de asociación que se pueden utilizar para calcular la significación de una coocurrencia léxica, así como otras medidas de distribución de unidades léxicas en un corpus. Algunos de estos coeficientes que presentaré primero son conocidos en la literatura, en algunos casos con más de cuarenta años de antigüedad, como los coeficientes que describe Herdan (1964). Estas técnicas se podrían implementar ya mismo en Terminus si aun se pretende utilizar este programa como extractor terminológico, y si bien son técnicas no demasiado efectivas, seguro superan la pretendida extracción mediante el simple conteo de palabras. Y la segunda parte es la propuesta de extractor que les traigo. Este algoritmo está basado también en estadísticas léxicas pero aprovechando la inmensa cantidad de datos que nos ofrece Internet, interrogando de manera automática a los motores de búsqueda comerciales. La estrategia es no sólo estudiar las frecuencias de las palabras en el texto que estamos analizando comparadas con las frecuencias que esas palabras registran en la web, sino también estudiar las diferentes relaciones que se producen entre las palabras tanto en el documento como en la web. Esta segunda parte del seminario está basada en una ponencia que presenté en el último congreso de AESLA (Nazar, 2008), aunque en ese caso el tema del artículo no era específicamente la extracción de terminología sino una aproximación a un problema muy antiguo en semántica que es la diferencia entre la utilización de un signo lingüístico con una función predicativa respecto a su utilización con una función referencial, o bien, siguiendo la terminología de Frege (1892), la distinción entre referencia y sentido. La novedad de aquella propuesta era plantear desde una formalización estadística, mediante la lingüística de corpus, un problema que pertenece tradicionalmente a la filosofía y a la semántica. Es, indirectamente, una consecuencia teórica de mi trabajo de tesis en curso y su aplicación práctica es la detección de unidades referenciales en un texto, conjunto que incluye al de las unidades terminológicas.

2. Algunos coeficientes de asociación y de distribución

En esta parte hablaré de coeficientes de asociación y distribución. Estas medidas dependen en todos los casos de la ayuda de un corpus de referencia del lenguaje general de la lengua en que está escrito el texto cuya terminología queremos extraer. La idea de usar un corpus de este tipo es poner en relación los datos de la frecuencia de aparición de las unidades analizadas respecto a este corpus de referencia que nos da la expectativa normal de aparición de una unidad. En la actualidad, conseguir un corpus de referencia de algunos millones de palabras es una tarea relativamente sencilla. Disponemos ya de corpus en distintas lenguas y también de las herramientas para compilar los de lenguas que aun no hemos estudiado. En este caso utilizamos como corpus de referencia el mismo corpus del IULA en la parte de lenguaje general del castellano.

2.1. Medidas de asociación

El propósito general de las medidas de asociación estadística es determinar si dos eventos están relacionados o si su concurrencia puede atribuirse al azar. Si dos eventos tienen lugar de manera independiente con una alta frecuencia, entonces es de esperar que se produzcan los dos al mismo tiempo sin que por ello tengamos que pensar que la ocurrencia de ambos está relacionada. Por el contrario, si observamos que dos eventos ocurren raramente por separado pero lo hacen con relativa frecuencia de manera conjunta, entonces sí vamos considerar probable que entre ambos exista una relación.

En el escenario de la extracción de terminología definimos un evento como la ocurrencia de una unidad léxica en un texto. Pero entonces, si estas son medidas de asociación, hay que decidir

primero cuáles son los elementos cuya asociación queremos estudiar. Si vamos a concentrarnos en la terminología polilexemática, entonces tal vez tenga sentido plantearnos el estudio de la asociación entre los distintos componentes de una combinación. Por ejemplo, si estudiamos un término como *asma bronquial*, calcular entonces la asociación entre *asma* y *bronquial*. Pero podríamos plantearlo de otra manera, como por ejemplo calcular la asociación existente entre un término y todos los que ocurren en el documento o corpus estudiado dentro de una ventana de contexto de *n* palabras. O bien, elegir un elemento cualquiera dentro del documento, a modo de término *pivot*, por ejemplo un elemento en el título del documento -o en cualquier parte-, y estudiar luego la asociación entre cada uno de los elementos del documento respecto a ese término.

Una forma fácil o intuitiva puede ser cuando la aplicamos al estudio de unidades polilexemáticas. En este caso podemos estudiar la fuerza de atracción entre los distintos miembros de un bigrama o cadena de palabras. En este caso entre *asma* y *bronquial*. Podemos hacer una prueba interrogando el corpus del IULA con el programa CQP, en primer lugar con este comando:

```
[word="asma"][word="bronquial"];
```

Esto es, le pedimos la cantidad de bigramas donde se realiza *asma bronquial*. La respuesta es 401.

```
[word="asma"][word!="bronquial"];
```

Este otro es la cantidad de bigramas donde se realiza asma pero no bronquial, es decir asma seguido de cualquier otra, y la respuesta es 769.

```
[word!="asma"][word="bronquial"];
```

Finalmente, le pedimos la cantidad de bigramas en los que el primer componente es cualquier cosa menos *asma* y el segundo *bronquial*. La respuesta es 299. En este momento el corpus del IULA en la parte de castellano tiene 23.505.935 tokens (que es la cantidad de bigramas + 1). Podemos plantear entonces esta relación en una tabla de contingencia de 2x2:

	w_1 = asma	w_1 != asma
w_2 = bronquial	401	299
w_2 != bronquial	769	23504465

Tabla 1: tabla de contingencia de 2x2 con los valores observados

En la celda superior izquierda tenemos indicada la cantidad de bigramas donde se realiza *asma bronquial*, la celda superior derecha la realización de *bronquial* sin *asma*, la inferior izquierda la de *asma* sin *bronquial*, y la inferior derecha la cantidad la cantidad total de bigramas menos aquellos en donde se realiza *asma* o *bronquial*. La fórmula de la chi-cuadrada que aparece más abajo expresa la suma del cuadrado de las diferencias entre los datos observados y los esperados en cada una de las celdas de la tabla, normalizadas por el valor esperado.

$$X^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Chi-square

No creo que sea necesario conocer al detalle su funcionamiento, en la actualidad es suficiente con presionar una tecla y las unidades aparecen listadas. De cualquier forma conviene tener una idea general de lo que hace y por qué se usa. O_{ij} serían los valores observados en cada celda de la tabla y E_{ij} los valores esperados. Con el propósito de conocer los valores esperados calculamos las frecuencias marginales (tabla 2), anotados en los márgenes de la tabla, y que solamente expresan los valores de las sumas de las filas y columnas respectivamente. La tabla 3 muestra cómo se calculan los valores esperados, multiplicando las frecuencias marginales que corresponden a cada celda, normalizado por el total de bigramas.

	w_1	w_1 !	
w_2	O_{11}=401	O_{12}=299	R_1=(401+299)
w_2 !	O_{21}=769	O_{22}=23504465	R_2=(769+23504465)
	C_1=(401+769)	C_2=(299+23504465)	N= 23505934

Tabla 2: Frecuencias marginales.

	w_1	W_1 !
W_2	$E11 = (R1\ C1) / N$	$E12 = (R1\ C2) / N$
W_2 !	$E21 = (R2\ C1) / N$	$E22 = (R2\ C2) / N$

Tabla 3: Valores esperados.

Reemplazando signos podemos reducir la expresión de la siguiente forma:

$$X^2 = \frac{N(O_{11}O_{22}-O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})}$$

El resultado en este caso es 4614678.3243. Este valor, en forma aislada, no es muy informativo. En realidad sólo lo necesitamos porque con él podemos asignarle al bigrama *asma bronquial* una posición en una lista de bigramas, es decir para presentar un ranking de bigramas en función de la significación estadística de su coocurrencia.

Otra medida bastante conocida entre lingüistas, en particular después de la aparición de un artículo de Church y Hanks (1991), es la información mutua, herencia de teoría de la información de Shannon (1948), expresada de la siguiente manera:

$$I(x,y) = \log_2 \frac{P(xy)}{P(x)P(y)}$$

Mutual Information
MI=4.6477 bits

P(xy) sería la probabilidad de ocurrencia de *asma bronquial*, *P(x)* la de *asma* y *P(y)* la de *bronquial*. Daille (1994) reporta una variante elevando el numerador al cubo. Según ella esto mejora los

resultados de esta medida cuando se la aplica a la extracción terminológica.

Como dije, la utilización de estas medidas, así como otras, mejora notablemente cuando disponemos de un corpus de referencia de lenguaje general con el cual podamos formar un modelo de la expectativa de ocurrencia normal de una unidad léxica. Como ejemplo presento un pequeño experimento, también con *asma*. Teniendo el corpus de referencia podemos hacer el experimento con un corpus no demasiado extenso. Utilizando este término como expresión de búsqueda, vamos a descargar las primeras 50 páginas web que devuelva alguno de los motores de búsqueda comerciales. Si ordenamos por frecuencia decreciente los bigramas donde aparece *asma* obtenemos la siguiente tabla:

nº	Unidad	Frec. Absoluta	Frec. Relativa
1)	asma ocupacional	35	0.00044121
2)	asma bronquial	32	0.00040339
3)	asma inducido	27	0.00034036
4)	asma laboral	13	0.00016388
5)	asma leve	12	0.00015127
6)	tienen asma	12	0.00015127
7)	tiene asma	9	0.00011345
8)	asma moderado	8	0.00010085
9)	asma asma	8	0.00010085
10)	asma severo	8	0.00010085
11)	asma infantil	8	0.00010085
12)	asma alérgico	7	0.00008824
13)	lâ asma	6	0.00007564
14)	asma persistente	6	0.00007564
15)	asma grave	5	0.00006303
16)	asma tienen	5	0.00006303
17)	□ asma	5	0.00006303
18)	asma agudo	4	0.00005042
19)	asma desde	4	0.00005042
20)	yo asma	4	0.00005042
21)	asma moderada	4	0.00005042
22)	asma incluyen	4	0.00005042
23)	asma inducida	4	0.00005042
24)	□ asma	4	0.00005042
25)	provocar asma	3	0.00003782
26)	asma episódica	3	0.00003782
27)	asma antes	3	0.00003782
28)	estilo asma	3	0.00003782
29)	asma causado	3	0.00003782
30)	asma m	3	0.00003782
31)	asma le	3	0.00003782
32)	asma design	3	0.00003782
33)	asma aguda	3	0.00003782
34)	padecen asma	3	0.00003782

Tabla 5: bigramas ordenados por frecuencia

Observamos que entre los bigramas hay muchos frecuentes que no son significativos, aunque apliquemos una *stoplist.* Si, en cambio, reordenamos esta lista por el valor que obtienen calculando la X^2 y tomando el corpus de referencia, observamos que entonces los bigramas aparecen un poco mejor ordenados, ya que los menos significativos tienden a aparecer en la parte inferior de la tabla.

nº	forma	frec	modelo Es	CHI
1	asma ocupacional	31	0	991.912.809
2	asma bronquial	26	0	826.952.693
3	asma inducido	25	8	793.969.563
4	lâ asma	6	0	169.850.013
5	asma alérgico	5	0	137.607.141
6	□ asma	4	0	105.679.049
7	□ asma	4	0	105.679.049
8	asma severo	8	23	87.973.129
9	asma design	3	0	74.301.807
10	asma episódica	3	0	74.301.807
11	asma leve	8	31	65.940.053
12	asma persistente	5	18	43.386.684
13	asma agudo	4	14	35.173.158
14	asma aguda	3	8	32.989.759
15	asma moderado	7	55	28.779.108
16	asma laboral	13	197	27.951.410
17	asma moderada	3	23	12.333.772
18	asma m	3	110	2.620.116
19	tienen asma	10	1212	2.529.255
20	asma infantil	3	152	1.884.638
21	estilo asma	3	220	1.286.659
22	asma grave	3	268	1.046.640
23	asma pueden	5	1035	700.878
24	asma puede	7	2855	436.048
25	tiene asma	7	3221	372.730
26	tener asma	3	1162	199.338
27	yo asma	4	2757	120.167
28	asma también	5	5119	76.977
29	asma qué	3	4175	23.886
30	asma son	3	4346	21.600
31	sobre asma	3	5449	11.080
32	asma le	3	7644	0.2040

Tabla 6: bigramas ordenados por la chi-cuadrada

Esto nos dice qué tan rara es una combinación de palabras y por lo tanto qué tan informativa es. Si hiciéramos lo mismo con un corpus más grande obtendríamos mayor cantidad de candidatos. Entre ellos vendrá, naturalmente, ruido, ya que la web es un corpus sucio, y por lo tanto tendremos que acostumbrarnos a ver símbolos no decodificados o diseñar algún filtro efectivo para depurarlo. El registro número 4, por ejemplo, aparece resaltado por su rareza en el corpus de referencia y su frecuencia en el corpus observado. Sin embargo esa rareza se debe a un carácter no reconocido.

El ordenamiento que obtenemos utilizando la información mutua es bastante similar:

nº	Forma	frec	modelo Es	MI
1	asma ocupacional	31	0	166.442
2	asma bronquial	26	0	163.904
3	lâ asma	6	0	142.749
4	asma alérgico	5	0	140.119
5	□ asma	4	0	136.900
6	□ asma	4	0	136.900
7	asma design	3	0	132.749
8	asma episódica	3	0	132.749

9	asma inducido	25	8	131.639
10	asma severo	8	23	101.050
11	asma aguda	3	8	101.050
12	asma agudo	4	14	97.831
13	asma persistente	5	18	97.640
14	asma leve	8	31	96.900
15	asma moderado	7	55	86.900
16	asma moderada	3	23	86.900
17	asma laboral	13	197	77.611
18	asma m	3	110	64.805
19	asma infantil	3	152	60.176
20	estilo asma	3	220	54.870
21	asma grave	3	268	52.035
22	tienen asma	10	1212	47.675
23	asma pueden	5	1035	39.951
24	tener asma	3	1162	30.913
25	asma puede	7	2855	30.176
26	tiene asma	7	3221	28.436
27	yo asma	4	2757	22.606
28	asma también	5	5119	16.900
29	asma qué	3	4175	12.470
30	asma son	3	4346	11.891
31	sobre asma	3	5449	0.8629
32	asma le	3	7644	0.3746

Tabla 7: bigramas ordenados por información mutua

2.2. Medidas de Distribución

Acabamos de estudiar la frecuencia de aparición de un término o de una combinación de términos en un documento y la pusimos en relación con su expectativa de aparición tomando como base de un modelo previo que teníamos de la lengua a partir del corpus de referencia. Ahora estudiaremos otra familia de medidas estadísticas para incluir también otro factor que es la forma en se distribuyen los términos en la colección analizada. Para eso podemos utilizar algunas medidas de dispersión ya reportadas (Sparck Jones, 1972) como TF.IDF (*term frequency x inverse document frequency*) que expresada de la siguiente forma:

$$w(i,\ j) = \left(1 + \log\ \left(tf_{i,j}\right)\right)\log\ \frac{n}{df_i}$$

Donde $tf_{i,j}$ es la frecuencia del término i en el documento j; df_i el número de documentos donde aparece i; cf_i como la frecuencia total de i en el corpus, n como el número total de documentos y suponiendo $tf_{i,j}$ es mayor que 0. Esta medida nos sirve para resaltar aquellos términos que tienen una frecuencia alta en un número limitado de documentos dentro de una colección. El supuesto es que los términos menos informativos tienen una frecuencia y una dispersión homogénea en la colección, es decir que su aparición no tiene relación con la temática del documento.

Para incluir también el corpus de referencia, que nos provee el factor que llamaremos Ef_i como la expectativa de aparición normal del término i, hemos implementado también una medida en el paquete estadístico *Jaguar*, que se expresa de la siguiente forma;

$$w_i = \log\left(\frac{cf_i \cdot df_i}{Ef_i}\right)$$

Índice de Jaguar

Esto nos permite trabajar de manera rápida sobre todos los términos del corpus, y tiene el efecto de resaltar aquellos términos que:

1) tienen una alta frecuencia en el corpus analizado;

2) están bien dispersos dentro de la colección analizada y

3) son raros en el corpus de referencia. Como ejemplo incluimos el listado extraído de las mismas 50 páginas que descargamos de Internet con *asma*.

Rank	forma	frec	docs	mod	peso
1	asma	916	36	1	29.624
2	inhalados	82	4	1	19.191
3	bronquial	78	10	1	18.976
4	alergenos	57	8	1	17.634
5	agonistas	46	4	1	16.721
6	corticosteroides	43	3	1	16.435
7	salmeterol	41	3	1	16.232
8	alergia	35	10	1	15.563
9	ocupacional	35	2	1	15.563
10	allergy	34	3	1	15.441
11	ácaros	33	5	1	15.315
12	asmáticos	33	6	1	15.315
13	asma ocupacional	31	2	1	15.051
14	inmunoterapia	30	3	1	14.914
15	disnea	30	3	1	14.914
16	beta	29	2	1	14.771
17	broncodilatadores	29	5	1	14.771
18	asma bronquial	28	6	1	14.624
19	alergias	28	4	1	14.624
20	prolongada	27	3	1	14.472
21	□Â □n	27	2	1	14.472
22	sibilancias	26	5	1	14.314
23	orales	25	3	1	14.150
24	alergeno	24	3	1	13.979
25	ción	22	3	1	13.617
26	desencadenantes	21	5	1	13.424
27	salbutamol	20	4	1	13.222
28	teofilina	20	3	1	13.222
29	obstrucción	19	5	1	13.010
30	bronquios	19	5	1	13.010
31	asmática	18	4	1	12.788
32	clin	18	2	1	12.788
33	respir	17	2	1	12.553
34	agonista	17	2	1	12.553
35	formoterol	17	3	1	12.553
36	corticosteroides inhalados	16	2	1	12.304
37	inhalado	16	2	1	12.304
38	placebo	15	3	1	12.041

39	cochrane	15	2	1	12.041
40	alérgica	14	5	1	11.761
41	inhalador	14	3	1	11.761
42	utilizarse	14	2	1	11.761
43	asmático	14	5	1	11.761
44	immunol	14	2	1	11.761
45	inhaled	13	2	1	11.461
46	miento	13	3	1	11.461
47	inmunología	13	2	1	11.461
48	asthma	98	5	8	11.222
49	espiratorio	12	4	1	11.139
50	expectoración	12	2	1	11.139

Tabla 8: Unidades ordenadas por el coeficiente de Jaguar

Otra ventaja que tiene esta medida es que ya no estamos estudiando la asociación que tiene un término con otro, y por lo tanto podemos trabajar de manera mucho más rápida porque ya no tenemos que hacer una sucesión de comparaciones sino simplemente ordenar todas las unidades en base a esta ponderación.

Pregunta de un miembro de la audiencia: *Pero este índice, esta ponderación final que tiene cada unidad en el listado, ¿qué quiere decir exactamente? Que asma aparece en mayor cantidad de documentos y que por tanto tiene una alta dispersión y por lo tanto tiene mayor valor terminológico que uno que aparece irregularmente?*

Sí, justamente. Pero es la combinación de factores, no sólo la dispersión. ¿Qué significa que un término tenga una baja ponderación? En este listado, los primeros 50 términos se pueden considerar valiosos desde el punto de vista terminológico, pero si continuáramos observando los rangos inferiores de la lista veríamos aparecer las unidades menos importantes. Una palabra que sea muy común, es decir que sea utilizada con regularidad en el corpus de referencia general de una lengua no va a ser muy beneficiada en este estudio, porque aquí estamos resaltando palabras que son frecuentes y dispersas en el corpus analizado y al mismo tiempo raras en el corpus de referencia.

Pregunta de otro miembro de la audiencia: *Todas estas unidades salen del corpus que estás analizando...*

Claro, las unidades que estamos estudiando, es decir, los candidatos a témino, salen del documento o corpus bajo análisis. En este caso sigue siendo el mismo corpus de las primeras 50 páginas web que bajé de internet con el término *asma* como expresión de búsqueda.

Pregunta de otro asistente: *¿Y estos documentos que descargas tienen algún tipo de limpieza o procesamiento? ¿No hay necesidad de eliminar de ese corpus aquellos documentos que no sean de medicina? ¿Es el corpus en bruto?*

No, limpiar significa simplemente eliminar todas las etiquetas html, pero no hay necesidad de ningún tipo de procesamiento. Es el corpus en bruto, exactamente.

Pregunta de otro asistente: *Pero el problema de estas medidas es que no serán efectivas en el caso de las unidades polisémicas, es decir, aquellas unidades que son formalmente idénticas a una que se utiliza en el lenguaje general, y que por lo tanto no es rara desde el punto de vista de estas medidas, y sin embargo sí es terminológica.*

Exactamente, ese es un problema grave y una limitación de este tipo de medidas. Es buena la observación. Creo que estas medidas, si se van a utilizar, tienen de bueno que son fáciles de implementar y de rápida ejecución. Pero no creo que sea una buena idea utilizarlas porque si bien, justamente como tú dices, pueden tener una alta precisión, es decir, puede que los términos que elija sean valiosos desde el punto de vista terminológico, no necesariamente van a tener una alta cobertura, porque van a dejar muchos términos en el documento sin analizar. Y sobre todo, la limitación más grave que tienen, es la dependencia con la frecuencia de aparición. Una palabra que aparece menos de tres veces en el corpus no puede ser analizada, y esto es un problema porque la frecuencia de aparición no tiene una correlación con su valor terminológico. Un documento puede tener cualquier cantidad de terminología entre sus hapax legomena. Entonces tenemos que buscar otro tipo de estadística, un método que nos permita hacer la estadística a partir de un solo caso, cosa que es, desde el punto de vista de la propuesta que les traigo ahora, efectivamente posible.

3. La propuesta del extracción terminológica (y su relación con la problemática de la distinción entre referencia y sentido)

Se plantea esto como un problema puramente matemático, es decir, sin conocimiento de la lengua y del dominio temático, porque esto es lo que da la versatilidad de una estrategia. Ahora, ¿cuál es la relación de esto con la problemática de la referencia y el sentido? La relación es con una ponencia que presenté en el último congreso de AESLA, (Nazar, 2008) aunque allí no hablaba específicamente de terminología, si no del problema de la referencia y el sentido. Sin embargo ahora quiero aplicar la misma metodología al estudio de la terminología especializada.

En lo que resta del seminario me voy a guiar, entonces, por ese artículo, pero solamente una parte porque allí planteo dos tipos de experimentos, uno desde el punto de vista diacrónico y el otro sincrónico. Ahora sólo comentaré el estudio desde el punto de vista sincrónico porque parece ser suficiente para la extracción de terminología.

El problema de la distinción entre referencia y sentido es conocido en lógica y en filosofía del lenguaje. El antecedente principal es sin duda el artículo *Über Sinn und Bedeutung* de Frege (1892), sin embargo hay un antecedente previo en el artículo *On a New List of Categories*, escrito por Peirce en sus primeros años (1867). Frege aparentemente no conocía a Peirce, sin embargo ambos artículos ofrecen algún paralelismo. Hay autores (Uxía Rivas, 1996) que ponen en relación directa el signo, referencia y sentido de Frege con la concepción triádica del signo de Peirce. En cualquier caso, la obra de Peirce ha sido históricamente ignorada en Europa en semántica (Heger, 1974; Lyons, 1980) y particularmente en el debate que despertó la obra de Frege (Russell, 1905; Strawson, 1950). Posiblemente esta ausencia pueda deberse a la desfavorable divulgación que hicieron Ogden y Richards (1923) de la concepción triádica de Peirce, tal como señala Eco (1968). El debate tiene todavía actualidad tanto en filosofía como en semántica formal (Putnam, 1975; Kamp, 1981; Katz, 2005). La mayoría de las veces se cita a Frege, excepto tal vez en los autores americanos, entre los que Peirce tiene mayor influencia. De cualquier forma no parece necesario entrar en detalles teniendo en cuenta que nuestro interés es bastante práctico, desde el punto de vista de la terminología.

De la literatura señalada vamos a extraer solamente algunos conceptos básicos: decimos que una expresión tiene referencia o valor referencial cuando designa un objeto definido. El ejemplo típico de tal expresión es el nombre propio. Podemos referirnos a un famoso filósofo mediante su nombre, *Aristóteles*, pero también mediante una descripción definida, como *el discípulo de Platón*, o también como *el maestro de Alejandro*. En su artículo Frege se pregunta por la relación de

equivalencia entre signos, como la que se da entre estos que refieren a Aristóteles, relación que algunos podrían llamar de sinonimia. Frege se pregunta específicamente qué significa que A sea igual a B. Decir que A es igual a A es, desde el punto de vista lógico (por lo menos el de la lógica del siglo XIX) algo trivial o poco informativo, ya que se trata del principio de identidad y no contradicción, según el cual un elemento no puede no ser igual a sí mismo. Ahora, si decimos que A es igual a B, en realidad lo que estamos diciendo es que se trata de dos signos que tienen el mismo referente. Si los signos fuesen exactamente equivalentes, entonces su intercambiabilidad en cualquier enunciado no sería problemática. Y sin embargo sabemos que esto no es así. Si la variable A tiene el valor "Scott" y B tiene el valor "el autor de Waverly", entonces, dos enunciados como 1) y 2) no transmiten, evidentemente, la misma información:

1)Scott es Scott.
2)Scott es el autor de Waverly.

Lo que Frege nos diría en este caso es que, a pesar de que los dos símbolos tienen un mismo referente (Bedeutung), tienen un distinto "modo de darse", es decir, distinto sentido (Sinn). Ambos modos de relacionarse los signos con sus significados tienen hasta cierto punto un funcionamiento independiente, ya que podemos tener enunciados específicos que son perfectamente aceptables desde el punto de vista gramatical y tienen un sentido y sin embargo carecen de referente, porque el enunciado está en función de las circunstancias de la enunciación. La aceptabilidad de estas expresiones no tiene entonces ninguna relación con la lengua sino que depende de la información que es transmitida por el texto y un estado del mundo (real o ficcional). En nuestros días, la expresión 3), por ejemplo, tiene sentido pero no tiene referencia.

3) El actual rey de Francia es calvo.

Vamos a dejar aquí la discusión acerca de la referencia y el sentido, para pasar a concentrarnos en la manera en que esta distinción puede, por un lado, hacerse de forma estadística, y por otro, en qué forma esta distinción nos puede ayudar a llevar a cabo la detección de terminología. El planteo aquí es específicamente que en un corpus determinado podríamos llegar a determinar cuáles son las unidades léxicas que cumplen una función referencial en el discurso, siempre y cuando ese referente exista como una unidad cultural, es decir que sea percibida socialmente. Esto último es fundamental porque estamos utilizando Internet como un corpus de referencia, y, si ese referente o concepto no existe en Internet, no tendremos manera de identificarlo.

La terminología especializada es como los nombres propios ya que su significado es su valor referencial (Wüster, 1979; Cabré, 1999). El problema que vamos a tener aquí es que no todas las unidades referenciales son unidades desde el punto de vista terminológico, un nombre propio no es tan interesante para el terminólogo, y el algoritmo que presento aquí todavía no se fija en esta distinción. Sin embargo, esta distinción entre un nombre propio y un término es una tarea relativamente sencilla de llevar a cabo con estrategias superficiales. Siempre y cuando no trabajemos con una lengua como el alemán, podemos apelar a algo tan sencillo como la probabilidad de que la unidad en cuestión aparezca con mayúsculas. Obviamente muchas palabras que no son nombres propios pueden comenzar con mayúsculas, sobre todo si aparecen al principio de una frase, pero la probabilidad es muy inferior si la comparamos con la de un nombre propio. Sin duda hay, además, muchas particularidades de los nombres propios respecto al resto de las unidades, por eso no creo que radique aquí el problema.

3.3. La hipótesis que guía el trabajo

La hipótesis es entonces la siguiente: podemos distinguir, dentro de un texto, los términos que tienen función predicativa (una función *de sentido*) y los que tienen función denominativa (o *referencial*) en el discurso. Esta distinción se puede hacer mediante estadísticas distribucionales de los términos de un corpus que nos ayudan a caracterizar el *vecindario léxico* típico de una unidad determinada, y poner entonces en relación ese vecindario típico con el contexto léxico en el cual encontramos la unidad en cuestión. Este solapamiento entre el contexto analizado y el contexto típico es el que estaría hipotéticamente correlacionado con la función referencial.

3.4. El algoritmo

Para explicar cómo funciona este algoritmo mostraré un pequeño experimento con un artículo de la revista Neurología, en el que el programa señala cuáles son las unidades que le parecen terminológicas. El artículo es de A. Alonso1, N. Egüés Olazábal y O. Ayo Martín. (2006) *Infección por virus de Epstein-Barr y esclerosis múltiple*, Neurología nº 21. Cito textualmente mi propio trabajo en esta parte:

> Dentro de este documento seleccionamos aleatoriamente un párrafo y marcamos en negrita los términos referenciales por medio de nuestro conocimiento de la lengua. En segundo lugar, marcamos en altas las unidades a las que corresponde una entrada en un diccionario terminológico del área (Mosby 2001), para introducir una medida objetiva. La idea es entonces desarrollar un algoritmo que sea capaz de llevar a cabo ese mismo marcado con un criterio puramente estadístico. Este algoritmo toma cada unidad del texto, siendo cada unidad una cadena de hasta tres palabras ortográficas siempre y cuando no comiencen o acaben como una palabra funcional, y genera un primer vector $\vec{t}$ con cada una de estas unidades:
>
> $$\vec{t} = (w_1, w_2, w_3 ... w_n)$$
>
> Los componentes obtienen una ponderación inicial (w_i) sobre la base de su rareza en un corpus de referencia de lengua general en castellano, de una extensión de dos millones de palabras. La ponderación es el logaritmo de la frecuencia del término en el texto analizado sobre la frecuencia que tiene en el corpus de referencia (f_i) más 1, por si no aparece. Se elimina toda unidad que tenga una frecuencia absoluta superior a 200 en el corpus de referencia, entonces, si $f_i < 200$:
>
> $$w_i = \log\left(\frac{t_i}{(f_i + 1)}\right)$$
>
> El siguiente paso es convertir a cada componente de $\vec{t}$ en un nuevo vector de términos ($t_i = \vec{i}$). Estos son términos estadísticamente asociados, pero no extraídos del corpus de referencia, sino de la web. Descargamos 100 documentos donde aparece el término candidato y calculamos la asociación ahora por Información Mutua, porque queremos estimar el grado de asociación entre dos unidades léxicas, por un lado $\vec{i}$, que es el término que da nombre al vector, y por otro lado cada uno de los componentes de $\vec{i}$, o sea i_j.
>
> $$MI(\vec{i}, i_j) = \log_2 \frac{P(\vec{i}\, i_j)}{P(\vec{i})P(i_j)}$$
>
> la relación entre cada unidad que estamos analizando, es decir cada candidato a témino, con cada una de las unidades que componen el corpus que descargado de internet con este

candidato como expresión de búsqueda.

Para la expectativa de la frecuencia normal de una palabra tomamos una vez más el corpus de referencia. Esto va a dar la ponderación de cada elemento de $\vec{i}$ para poder eliminar todos los componentes que tengan una información mutua inferior a 9 bits. A continuación, calculamos el solapamiento entre cada vector $\vec{i}$ con $\vec{t}$ recordando que el primero representa a cada candidato y el segundo el texto analizado. En este caso hemos elegido por comodidad sólo un pequeño fragmento del documento, pero para calcular el solapamiento léxico deberíamos incluir todo el documento analizado, o por lo menos el léxico que tenga una ponderación (w_i) superior a -1. Este es el solapamiento (O_{it}) entre $\vec{i}$ y $\vec{t}$, que acusa la cantidad de unidades compartidas en relación a las unidades totales en el texto analizado.

$$O_{it} = \frac{|\vec{i} \cap \vec{t}|}{|\vec{t}|}$$

En este experimento aceptamos un candidato como referencial si el solapamiento supera un umbral de 0.05.

3.5. Resultados preliminares

Las muestras 1 y 2 comparan los marcas hechas por el algoritmo, el diccionario y el informante. Las marcas hechas en negrita en la muestra 1 son las que hice yo antes del experimento. En altas, en la misma muestra 1, están las palabras a las que corresponde una entrada en un diccionario terminológico. En la muestra 2 están los términos señalados por el algoritmo, que utiliza distintos juegos de corchetes para expresar que hay términos dentro de términos. Por ejemplo, cuando dice *[[seroprevalencia] de [anticuerpos]]* quiere decir que ha encontrado: *seroprevalencia*, *anticuerpos* y *seroprevalencia de anticuerpos.*

En los 3 primeros años de vida y la **SEROPREVALENCIA de anticuerpos** frente al **VIRUS** es del 100% en la primera década. En estos casos la **infección** por **VEB** normalmente es **asintomática**. Por contra, en países desarrollados la mitad de los niños son **SERONEGATIVOS** cuando alcanzan los 10 años, adquiriendo la **infección** muchos de ellos durante la ADOLESCENCIA o la juventud. La mitad de estas **infecciones** tardías darán **síntomas** y las más severas se manifestarán como **MONONUCLEOSIS INFECCIOSA**. Esta **ENFERMEDAD** se caracteriza por la presencia de **FIEBRE**, **adenopatías** y **FARINGITIS**. Cerca del 5% de la población adulta de países desarrollados permanece sin INFECTAR a lo largo de su vida.

Muestra 1: términos marcados como referenciales por el informante (negrita) y por el diccionario terminológico (altas).

En los 3 primeros años de vida y la [[**seroprevalencia] de [anticuerpos]]** frente al virus es del 100% en la primera década. En estos casos la [[**infección] por [VEB]]** normalmente es [**asintomática]**. Por contra, en países desarrollados la mitad de los niños son [**seronegativos]** cuando alcanzan los 10 años, adquiriendo la [**infección]** muchos de ellos durante la [**adolescencia o la juventud]**. La mitad de estas [**infecciones tardías] [darán [síntomas]]** y las más severas se manifestarán como [[**mononucleosis] infecciosa]**. Esta [**enfermedad se caracteriza]** por la presencia de fiebre, [**adenopatías]** y [**faringitis]**. Cerca del 5% de la población adulta de países desarrollados permanece sin infectar a lo largo de su vida.

Muestra 2: términos marcados como referenciales por el algoritmo.

4. Conclusiones

En primer lugar hay que decir que esto es poco texto para evaluar. Hace falta una evaluación más intensiva. Pero a simple vista se ve que el desempeño no es malo, aunque hay errores y omisiones. Por ejemplo *enfermedad se caracteriza*. Hubiéramos preferido que no señale ésta, pero es probable que sea una combinación muy común en artículos de este tipo. También *darán un síntoma*. Por otro lado falta *infectar*, falta *fiebre*. *Fiebre* tal vez aparece en una diversidad de contextos muy grande. De todas formas parece un resultado prometedor.

Una explicación de los motivos por los cuales una técnica como esta puede dar resultados positivos es que estamos observando la consecuencia de un mecanismo que opera a nivel discursivo y que podríamos denominar, usando un término de la antigua retórica, como el exordio, que se representa en este caso en forma de redundancia. Cuando el autor introduce un término referencial en el discurso, se espera que éste sea presentado o definido respetando los principios vigentes en una comunidad, incluso aunque luego se predique algo distinto acerca del referente. Esto sucede en el caso de que el autor lo crea conveniente o necesario, es decir según su "modelo de lector" (Eco, 1979). Cuando un autor introduce un término en el discurso, tiene que hacerlo "acolchado de redundancia" (Bougnox, 1985).

Esto tiene como consecuencia la asociación estadística entre términos, la consecuencia de la acción de muchos autores individuales, que puede expresarse como una red en la que estos se van tejiendo y enlazando, en una estructura colectiva producto de una serie de conductas individuales.

El trabajo pendiente ahora es replicar el experimento con distintos artículos y consultar a especialistas acerca del desempeño. La utilización de diccionarios terminológicos como una forma de control automático de los resultados se ha revelado un método de evaluación poco fiable, y la consulta manual utilizando diccionarios resulta demasiado costosa para aplicarla a una serie de artículos. Está pendiente además encontrar un sistema de evaluación para un extractor terminológico. También podría solicitarse la ayuda de estudiantes para la evaluación, tal como ayudaron a evaluar el programa Terminus.

Finalmente, y como advertencia, en esta metodología hay problemas de dos tipos. El primero es de orden práctico, y es la lentitud en la respuesta que viene dada por el tiempo necesario para la descarga de documentos de Internet. Este problema difícilmente podría resolverse teniendo un corpus propio, ya que el tamaño que este corpus tendría que tener obligaría a disponer de una infraestructura para el manejo de corpus que no podríamos tener en el corto plazo, lo que nos obliga

a utilizar los servicios de buscadores comerciales, y este es el otro problema que surge relacionado con esto, un problema de tipo legal ya que surge cierta incógnita acerca de la legalidad de una acción semejante. Si se trata de un experimento científico no puede ser demasiado grave, pero no está del todo claro que podamos distribuir libremente programas que realicen automáticamente extracción de información desde terceras partes. Es un problema que no se si me compete a mí resolverlo o si tenemos que consultar a algún servicio jurídico.

Pregunta de un miembro de la audiencia: *¿Cuál es el tamaño mínimo de un documento para poder llevar a cabo este análisis?*

El tamaño no es tan importante en este caso, puede ser una sola página (ideal sería el tamaño de un artículo científico, que sea mucho más corto o mucho más grande perjudicará el resultado). Esto es así siempre y cuando se tenga el corpus de referencia. El corpus de referencia usado aquí tiene 3 millones de tokens. Es muy poco, y mientras más grande sea ese corpus, mejor. Pero creo que hay una cota mínima entre los 2 y 3 millones.

Pregunta: *Otra cosa, este marcaje que hace el algoritmo de cada término, ¿se puede ordenar en una tabla de acuerdo a su ponderación como términos?*

Comentario de otro asistente: *Claro, cada uno tiene asociado un puntaje. Hay algunos que son más probables y otros menos probables. El problema está en términos como fiebre, que son de una mayor generalidad.*

Sí, como fiebre aparece en una mayor diversidad de contextos, pareciera que tiene menos información.

4. REFERENCIAS

Bougnoux, D. (1998). “Introduction aus sciences de la communication”, Paris, La Découverte.

Cabré, T. (1999). “La Terminología: Representación y Comunicación”. Barcelona: Institut Universitari de Lingüística Aplicada.

Church, K., and Hanks, P., (1991); “Word Association Norms, Mutual Information and Lexicography”, Computational Linguistics, Vol 16:1, pp. 22-29.

Daille, B. (1994). “Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques”. Thèse de Doctorat en Informatique Fondamentale. Université Paris 7.

Eco, U. (1968). “La estructura ausente”. Madrid: Lumen.

Eco, U. (1981). “Lector in fabula la cooperación interpretativa en el texto narrativo”. Trad. de Ricardo Pochtar, Barcelona, Lumen.

Frege, G. (1892/1993). “On sense and reference”, in A.W. Moore (ed.) Meaning and Reference. Oxford: Oxford University Press.

Heger, K. (1974). “Teoría semántica :Hacia una semántica moderna”. Madrid, Alcalá.

Kant, I. (1781/1978). “Crítica de la razón pura”. Traducción de Pedro Ribas. Madrid: Alfaguara.

Kamp, H. (1981 / 2002). “A theory of truth and semantic representation”. En J. Groenendijk, Th. Janssen, y M. Stokhof, editores, “Formal Methods in the Study of Language”, pages 277 -- 322. Mathematisch Centrum Tracts, Amsterdam, 1981. Reproducido en en “Formal Semantics: The Essential Readings”, editado por Paul Portner y Barbara Partee, Oxford: Blackwell.

Katz, J. (2004). “Sense, reference, and philosophy”, New York, Oxford University Press.

Lyons, J. (1980). “Semántica”, Barcelona, Teide .

Mosby (2001). “Diccionario Mosby de Medicina, Enfermería y Ciencias de la salud”. Quinta edición. Madrid: Harcourt. Versión en lengua española de la 5.ª edición de la obra original en inglés: Mosby's

Medical, Nursing, and Allied Health Dictionary, Mosby Year Book, Inc.
Nazar, R.; (2008). “Diferencias cuantitativas entre referencia y sentido”, Actas del XXVI Congreso de AESLA (Asociación española de Lingüística Aplicada) Universidad de Almería - del 3 al 5 de Abril del 2008 .
Peirce, C. S. (1867). “On a New List of Categories”, Proceedings of the American Academy of Arts and Sciences 7 (1868), 287–298.
Putnam, H. (1975/1985). “The meaning of 'meaning'”. *Philosophical Papers*, *Vol. 2: Mind, Language and Reality*. Cambridge University Press
Quine, W. (1951). “Two Dogmas of Empiricism”. *The Philosophical Review* 60 (1951): 20-43.
Russell, B. (1905). “On Denoting”, *Mind,* (14) : 479-493.
Shannon, C. E. (1948) “A mathematical theory of communication”. Bell System Technical Journal, vol. 27, pp. 379-423 // 623-656.
Sparck Jones, K. (1972). “A statistical interpretation of term specificity and its application in retrieval”. *Journal of Documentation*, 28 (1): 11-21.
Strawson, P.F. (1950). “On Referring”, *Mind, (*235): 320-344.
Uxía Rivas, M. (1996) “Frege y Peirce: en torno al signo y su fundamento”, Anuario filosófico, Vol.: 29, Número: 56, pp.1211-1224
Vivaldi, J. (2001/2004). “Extracción de candidatos a término mediante combinación de estrategias heterogéneas”, Barcelona: IULA, Sèrie Tesis 9.
Wüster, E. (1979/1998). “Introducción a la teoría general de la terminología y a la lexicografía terminológica”, Barcelona: IULA, Sèrie Monografies 1.