Seminario IULATerm - Universidad Pompeu Fabra
Fecha: 09/02/2007

Título:
Explotación estadística de corpus: análisis conceptual y clasificación de documentos.
Dictado por Rogelio Nazar

Contenidos:

Introducción

Esta es una presentación de tres líneas de investigación que se encuentran en el marco de los métodos de análisis cuantitativo en lingüística. Tiene el objeto de demostrar la utilidad que unos conocimientos básicos de estadística pueden ofrecer para el estudio del lenguaje.

Algunos de los métodos presentados no son en absoluto nuevos en términos históricos, sin embargo es sólo en épocas recientes cuando parecen ganar mayor difusión. Sin duda uno de los factores que impulsan este desarrollo es la revolución tecnológica de las últimas décadas. Con la disponibilidad de poderosos recursos informáticos y la abundancia de material en Internet, al lingüista se le ofrecen unas posibilidades de análisis nunca vistas.

Esta presentación no exige conocimientos previos en la materia y ofrece algunas alternativas de análisis y herramientas informáticas de fácil utilización, que se encuentran actualmente en desarrollo en el IULA, pero que tienen prototipos públicamente accesibles y ejecutables a través de Internet.

La primera parte ofrece una visión general de conceptos y técnicas aplicables al análisis del texto mediante un software que por ahora se llama Jaguar, el cual permite usar el Corpus del Iula o cualquier otro corpus y explotarlo con extracción de enegramas, medidas de asociación, medidas de distribución y medidas de similitud. La segunda parte se concentra en las conclusiones que podemos sacar en el nivel conceptual a partir del estudio de la coocurrencia léxica en el corpus. Esta es la línea de investigación de mi tesis doctoral, dirigida por Jorge Vivaldi y Leo Wanner, cuya aplicación sería la extracción de mapas conceptuales. Finalmente hay una línea de investigación en clasificación automática de documentos con un algoritmo de aprendizaje supervisado. En principio se aplicó para clasificar textos por tema pero en el ForensicLab también ha sido utilizado con éxito como herramienta de atribución de autoría.

Antecedentes

Es difícil trazar un recorrido certero en el campo que podríamos llamar lingüística cuantitativa porque la literatura es extensa. No es una corriente de pensamiento específica, pero los trabajos que reúne tienen características diferentes a otras aplicaciones del pensamiento lógico-matemático en lingüística, como las gramáticas generativas, que también alcanzan un alto grado de rigor en sus exposiciones. Las

aplicaciones de la estadística a la lingüística se caracterizan por un marcado empirismo, frente al racionalismo introspectivo de los formalismos chomskianos.

Como referencias elementales mencionaré, como pioneros del campo y figuras prominentes, a George Kinsley Zipf (1949); posteriormente a Benoît Mandelbrot (1961), quien ejerció una influencia importante en la comunidad científica desde su posición en la IBM a partir de los años 60. En esta lista sin embargo faltaría la obra de Andrei Markov a principios del siglo XX y por otra parte la tradición de la teoría de la información iniciada por Claude Shannon (1948). El lector español encontrará una introducción a la historia de la lingüística cuantitativa hasta la década del setenta en Müller (1973) y en Marcus et al. (1978). Desde los años 80 la difusión de la informática permite a cada vez más lingüistas aplicar técnicas estadísticas para la explotación de corpus y la lexicografía experimenta una revolución teórica. Algunos ejemplos de esta corriente son los trabajos de Church y Hanks (1991); Sinclair (1991); Kilgarriff (1997); Williams (1998), entre otros. Más allá de la lexicología, la estadística se aplica en general en el procesamiento del lenguaje natural. Una referencia que cubre buena parte del desarrollo que tuvo este campo a lo largo de los años '90 es el manual de Manning y Shuetze (1999). A partir de aquí la lista de autores y teorías relacionadas se dispersa de manera que dificulta una exposición resumida que haga justicia, pero una publicación de referencia puede ser el Journal of Quantitative Linguistics[1].

Hacer abstracción de la lengua como objeto matemático implica un cambio radical en el punto de vista y por tanto una nueva serie de posibilidades y soluciones, por lo que la modelación matemática de la lengua no debe ser tenida en cuenta simplemente como una cuestión utilitaria. De cualquier modo las aplicaciones son muchas y muy diversas. Algunas ampliamente extendidas en lingüística son la lexicometría; el etiquetado morfológico; la extracción de terminología; de colocaciones; la alineación de corpus paralelos; optimización de sistemas de recuperación de información; extracción de léxico bilingüe; clasificación de documentos; etc.

[1] http://www.informatik.uni-trier.de/~ley/db/journals/jql/index.html

El proyecto Jaguar

Jaguar es un programa que ofrece algunas posibilidades de explotación estadística de corpus y surge a raíz del interés del IULA por explorar este campo. Se comenzó a desarrollar en mayo de 2006, y en julio de ese año se cerró una primera etapa y desde entonces no ha habido grandes cambios. Lo que hay hoy es un paquete con documentación, software y código fuente, accesible en Internet en la dirección http://brangaene.upf.es/proves/jaguar

El motivo por el que se implementó en primera instancia como una aplicación web y no un programa ejecutable en el cliente es que esto permite al programa explotar el Corpus Técnico del IULA, además de otras ventajas propias de las aplicaciones web, como ser accesible desde cualquier parte, ejecutarse en cualquier plataforma y que no haya necesidad de instalarlo. Sin embargo, si el corpus con el que se trabaja es muy grande (20Mb) sería más lógico implementarlo como un programa local, por la disponibilidad de memoria y para evitar el tiempo de transferencia cada vez que se ejecuta alguna función.

Está escrito en Perl, un lenguaje extendido en lingüística computacional, lo que facilita a un investigador integrar el código fuente de las funciones de este programa dentro de sus propios proyectos.

Es preciso advertir que, estando la herramienta en desarrollo, posiblemente sufrirá una metamorfosis en el futuro, incluyendo nuevas funciones y completando las que ya tiene.

Funciones del programa

En el estado actual, el programa presenta las siguientes funciones:

- Tratamiento de un corpus propio o del CT-IULA
- kwic (keyword in context)
- enegramas
- media, varianza, desviación estándar
- medidas de asociación
- medidas de distribución
- medidas de similitud

Corpus

La primer función del programa consiste en dar al usuario la posibilidad de trabajar con un corpus propio. La característica que tiene que tener este corpus es que sea un archivo de texto. Pueden subirse varios archivos, los cuales se acumulan en un solo corpus. El

tamaño máximo de una colección probado hasta el momento es un archivo de 40 megabytes, pero esto complica el procesamiento y alarga el tiempo en varios minutos.

El programa reconoce etiquetas de lematización y morfología, siempre y cuando sean compatibles con las que usa el Corpus del IULA. Si se interroga a este corpus, que ya está etiquetado, es posible hacerlo recurriendo a la rica expresividad de las reglas del programa CQP.

Kwic (keyword in context)

La interrogación de tipo KWIC consiste en que el usuario proporciona una o más formas textuales o expresiones regulares y el programa devuelve concordancias (contextos de un tamaño de palabras a izquierda y derecha definido por el usuario) en que esa forma textual se realiza. Esta es la función más básica y es similar a la que ya ofrece otra aplicación específica para esto, el software bwanaNet.

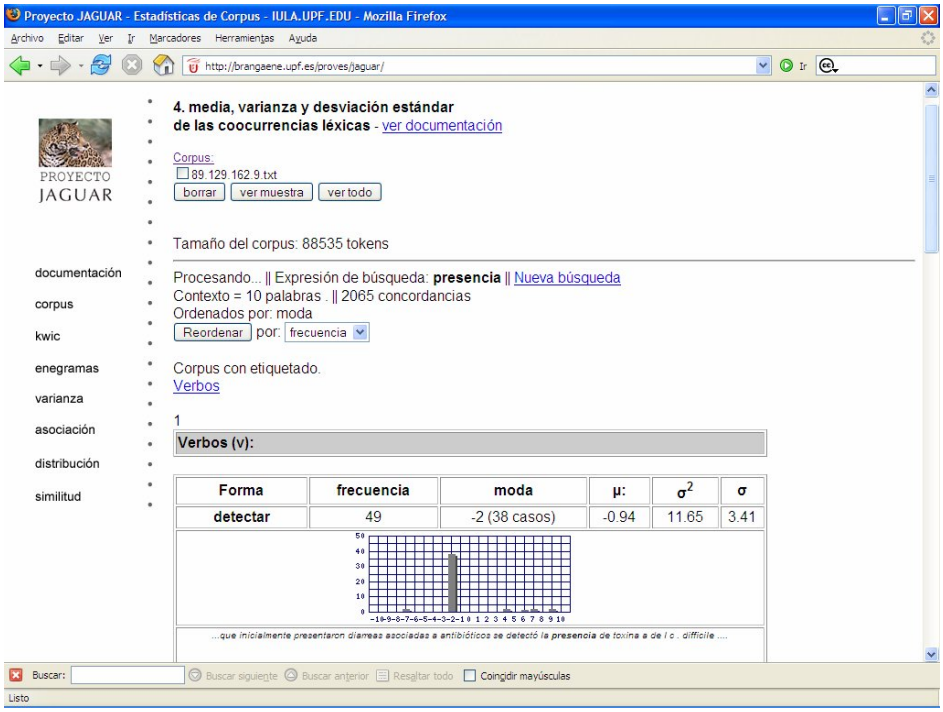
Enegramas

Esta función consiste en convertir un corpus en un vector con los enegramas de ese corpus (donde *n* puede variar de 1 a 5) ordenados en forma decreciente según su frecuencia de aparición. La literatura refiere este procedimiento como uno de los métodos más sencillos para descubrir colocaciones, nombres compuestos o frases terminológicas. También puede verse como una forma de medir la riqueza léxica, en el porcentaje de hapax o la relación entre el vocabulario y la extensión del texto.

Media, varianza, desviación estándar

En esta sección el programa toma muestras del corpus en las que se realiza una unidad léxica introducida por el usuario y devuelve histogramas donde se caracteriza el comportamiento de las unidades coocurrentes dentro de una ventana de contexto de un número de palabras definido por el usuario. Distintos parámetros, pero fundamentalmente la varianza de la muestra, permiten observar patrones de coocurrencia que no podrían observarse fácilmente de otra manera, por ejemplo cuando únicamente se estudia la frecuencia de aparición.

Si el corpus tiene etiquetas de categoría morfológica, es posible discriminar qué verbos acompañan típicamente a qué nombres, o adjetivos, etc. En la captura de pantalla vemos algunos histogramas pertenecientes a los verbos que más acompañan al sustantivo *presencia*. El verbo *detectar*, por ejemplo, aparece siempre dos posiciones antes del sustantivo, como en *detectar la presencia*. De esta manera los histogramas nos muestran que una presencia se puede *detectar*, *determinar* o *confirmar*, y también puede *aumentar*. Se puede *contar con la presencia* de alguien, una presencia se puede *relacionar* con algo o puede *indicar* o ser *indicada* por algo. Estos verbos surgen claramente como satélites del sustantivo, frente a otros, igualmente frecuentes pero no significativos por su distribución uniforme.



Medidas de asociación

Esta función es similar a Enegramas en cuanto presenta una lista de bigramas, pero esta vez no ordenados en base a su frecuencia de aparición sino a la asociación estadística entre los dos componentes. Las medidas utilizadas son t-test, chi-square y Mutual Information.

Medidas de distribución

En este apartado se analiza la distribución de frecuencias de un término, tanto dentro de un documento como dentro de un corpus. Este tipo de medidas puede ser útil para valorar la relevancia de un término en un documento. La frecuencia de aparición es una fuente de información, pero también podemos analizar la forma como se distribuyen las ocurrencias de ese término en el documento y, por otro lado, la distribución del término en toda la colección de documentos.

Si un término aparece solamente al principio de un documento, esto puede sugerir que el término solo se utiliza en la introducción, pero que no forma parte de la temática central del texto. Si, por ejemplo, utilizamos esta función para analizar este mismo texto, comprobaremos que las apariciones de la expresión *mapa conceptual* en su forma plural y singular, se producen sólo en el tercio central, que es donde se trata el tema.

Medidas de similitud

Esta función calcula medidas de asociación entre vectores, y está actualmente implementada para comparar expresiones de una o más palabras o textos completos. El programa convierte el corpus en trigramas de caracteres en el caso de las expresiones y enegramas de palabras (donde *n* es un parámetro definido por el usuario) cuando compara textos. Por el momento los coeficientes que se pueden utilizar son Matching; Jaccard; Dice; Overlap y Coseno.

Las medidas de similitud entre expresiones serían útiles, por ejemplo, si quisiéramos estudiar casos de variación terminológica o bien para encontrar documentos donde aparecen términos parecidos a nuestra expresión de búsqueda, como *psiquiátrico* y *siquiátrico* (dice 0.947); *heterozigoto* y *heterocigoto* (0.700); pero también *heterocigoto* y *homocigoto* (0.556). Sin embargo en una aplicación así estará el ruido de asociaciones como *cirrosis* con *virosis*, sólo porque tienen en común trigramas como "ros", "osi" y "sis".

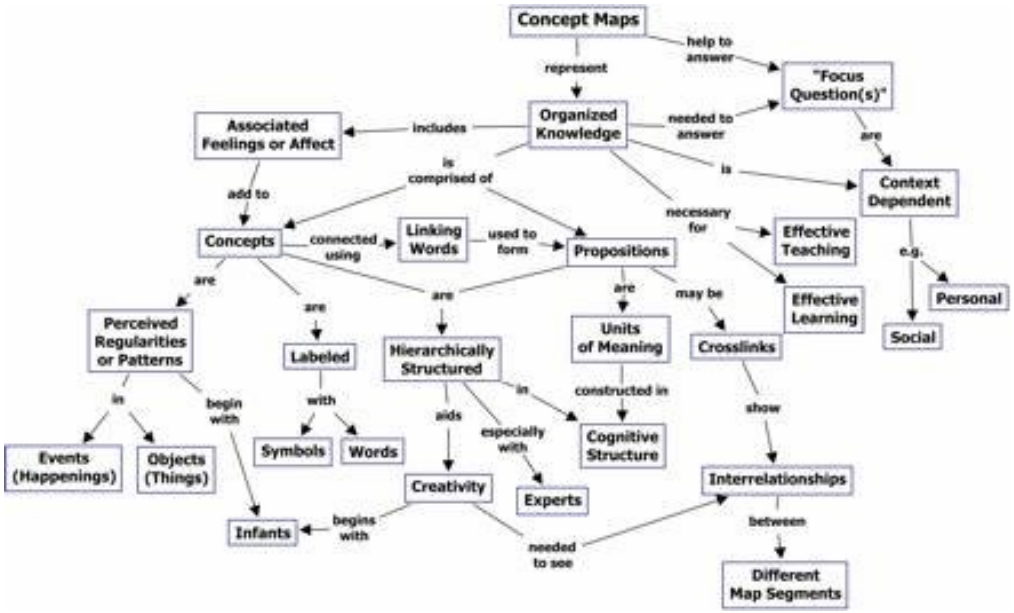
En cuanto a la similitud entre documentos, la manera como está implementada ahora da la posibilidad de obtener, a partir de un conjunto de documentos, un ordenamiento de esos documentos en base a la similitud que tengan con un documento que se le señale previamente como referencia. Sería útil para quien tenga un documento y desee conseguir documentos similares en el interior de una colección desordenada.

Análisis conceptual

La visión estadística del texto permite también un análisis conceptual. La estadística señala cuáles son los términos que tienen una tendencia a ocurrir juntos -a distancias variables el uno del otro dentro de un contexto reducido de texto- con una frecuencia superior a la que deberían tener si aparecieran así por casualidad. Esta es una regularidad matemática, aplicable a cualquier lengua actual o potencial, que rige la distribución del léxico en el discurso.

Esto permite que el sentido sea modelado algorítmicamente y expresado en una gráfica con las construcciones sintagmáticas más significativas. Metafóricamente, sería como fotografiar un concepto a partir de un corpus. Planteado como una función, requiere una entrada, que sería un término, y ofrece como salida un mapa conceptual.

Un mapa de estas características presenta similitudes con los mapas conceptuales que describe Novak (2001): una red de arcos y nodos donde los nodos están etiquetados con términos que representan entidades o conceptos y los arcos con predicaciones que indican relaciones conceptuales entre los nodos.



El mapa conceptual de los mapas conceptuales, dibujado por Novak.

La idea de Novak no tiene relación con algoritmos de recuperación de información, sino con una metodología de apoyo didáctico, en la que se supone que un estudiante debe dibujar manualmente mapas conceptuales que sintetizan lo que ha comprendido luego de la lectura de los textos de clase.

La propuesta de la tesis entonces es elaborar un algoritmo que haga en forma mecánica el

proceso de generación del mapa conceptual luego de la lectura de un conjunto de documentos.

Una de las aplicaciones posibles de este análisis sería un sistema de recuperación de información que presente al usuario un mapa conceptual de la consulta que ha hecho, mediante la extracción de una red de términos que representan entidades que están conceptualmente relacionadas con su consulta.

Si se aplicara como motor de búsqueda representaría una ventaja sobre los sistemas que se limitan a presentar como resultado de una búsqueda una lista de documentos, como lo hacen actualmente Yahoo y Google. Transformar los documentos, que están en una sola dimensión, a un mapa de dos o tres dimensiones, ayuda a poner orden en una colección de texto no estructurada, presentando un cuadro sinóptico con la información más relevante.

Hipótesis principal

Los términos tienen más o menos relación semántica en función de la frecuencia con la que aparezcan juntos -en grandes colecciones de documentos- en un contexto reducido de palabras. A partir de esa medida sería posible reconstruir el conjunto relaciones semánticas entre términos, según la idea bastante simple de que las proposiciones esenciales acerca de un objeto presentan una asociación estadística, en base a su frecuencia de co-aparición en el discurso, en comparación con las accidentales o circunstanciales. Si se proyecta el discurso, que está en un código serial, a un espacio de dos dimensiones, y establecemos las distancias entre los términos en base a su frecuencia de coocurrencia, observamos que los términos que hacen referencia a conceptos o entidades relacionadas tienen tendencia a aglutinarse como átomos en estructuras moleculares.

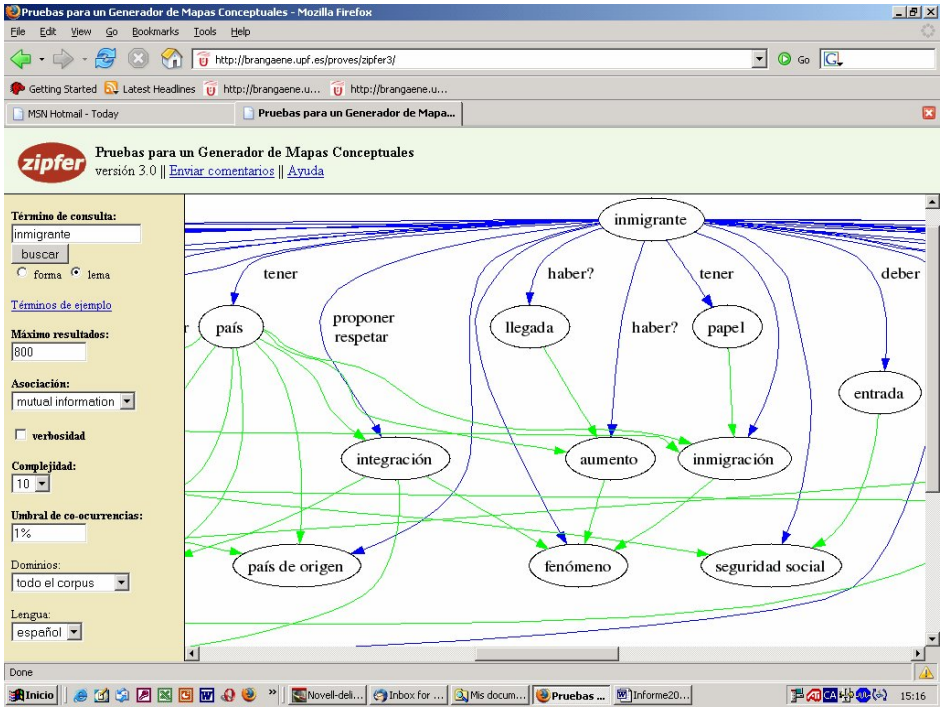
Este comportamiento de las unidades léxicas es también la forma de medir su valor referencial o su valor especializado, como una magnitud de información que puede estudiarse con métodos cuantitativos. Algunas palabras se distribuyen homogéneamente con una frecuencia regular en el discurso, en el extremo opuesto encontramos las unidades referenciales y las terminológicas, aquellas que tienen por referente un objeto del mundo real o imaginario, sea una entidad, un lugar, un proceso, una clase, etc. Podríamos incluir aquí a la terminología científica y a los nombres propios.

Experimentos

El procedimiento es sencillo, no se pretende que el algoritmo alcance una interpretación racional de los datos. Lo que hacemos es solamente estudiar la redundancia, y luego, en base al etiquetado morfológico, una distinción entre verbos para los arcos y sustantivos o determinadas frases nominales para los nodos. El gráfico resume cuáles son las

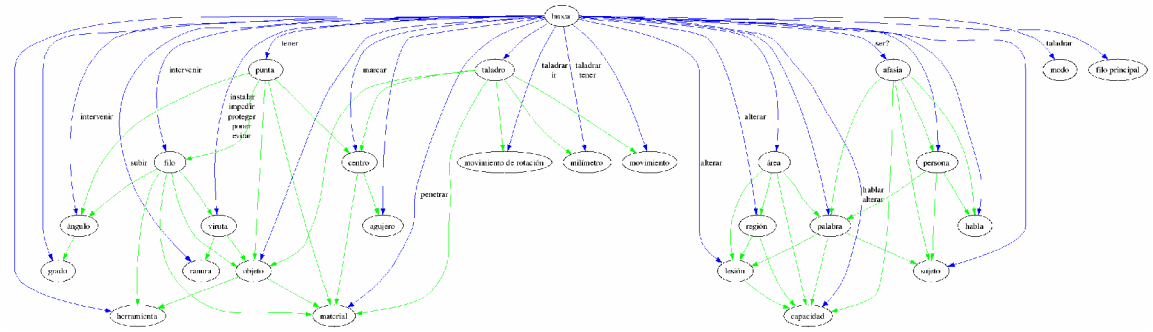
combinaciones más frecuentes, pero todavía no ofrece información sobre relaciones de dependencia. Esa es la parte del trabajo actual, y posiblemente sea haga con la ayuda de analizadores de dependencias ya existentes (como los de Atserias y otros, 2005; Attardi, 2006; Calvo y Gelbukh, 2006; entre otros posibles.)

El trabajo realizado hasta ahora es la captura de los términos que demuestren una asociación estadística significativa, usando medidas como chi-square y Mutual Information. Lo que hay hoy es un "esqueleto" conceptual, al que progresivamente se le adosará el tejido, en forma de capas más finas de significado. Gracias a la selección de términos relevantes de manera numérica, nos evitamos caer en el trampa de intentar analizar la lengua a través de la propia lengua. Así, no es necesario tener que explicitarle al algoritmo que una *casa blanca* es una cosa y otra muy distinta es *la Casa Blanca*, porque sólo la última tiene una referencia específica, que en el año 2006 selecciona un conjunto definido y prominente de términos-vecinos frecuentes como *Presidente Bush*, *Washington*, *EEUU*, *país*, *guerra*, etc.



Primeros esqueletos de un mapa conceptual.
http://brangaene.upf.es/proves/zipfer3

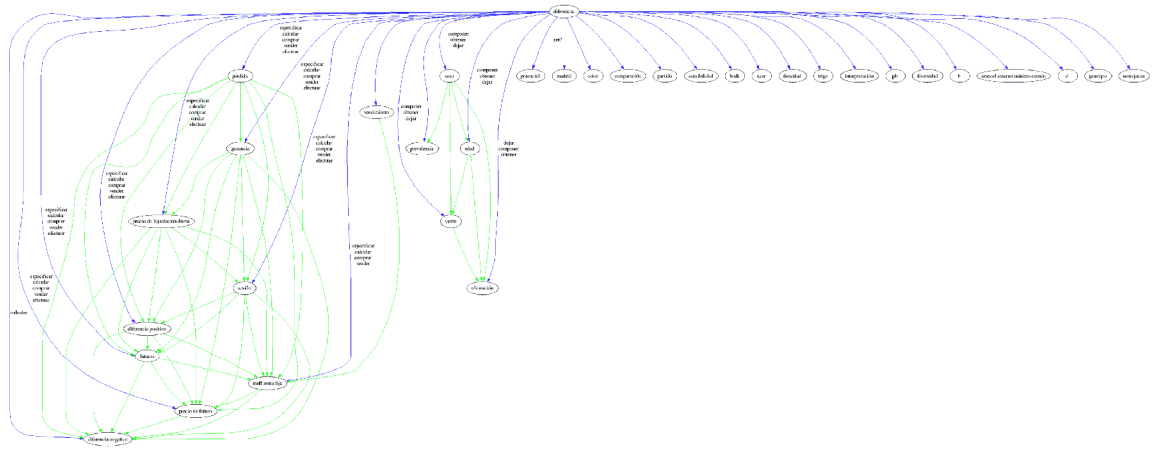
La referencia de un término se ve acusada por la de sus vecinos. Esto se ve más claramente cuando observamos las constelaciones que producen los términos polisémicos. En el siguiente ejemplo está el término *broca*, que en el Corpus del IULA es una región del cerebro y también la pieza de un taladro. Se advierte la formación de dos *clusters* agrupando los vecinos frecuentes de cada uno de los dos sentidos.



Aspecto del mapa conceptual de "broca".

Se podría decir que Broca es también un nombre propio, y que es un sentido que no está reflejado en el gráfico, pero ese es un uso que no se da en el CT-IULA, y este sistema sólo va a reflejar lo que hay en el corpus.

Otro ejemplo interesante por su polisemia es *diferencia*, que, aunque en un primer momento se pueda pensar que no tiene un uso técnico referencial, al observar su constelación descubrimos que sí lo tiene, al ser un término de la economía (diferencia es, después de todo, un término matemático, el opuesto a la adición) y forma un cluster con *diferencia positiva* y *diferencia negativa*, junto a *futuros*, *precio de futuro*, *precio de liquidación diaria*, etc.



Un ejemplo de vocablo no terminológico sería *importancia*: ahora sí, no es capaz de agrupar nada, no tiene referencia. Es una predicación que exige por lo menos uno o dos argumentos para tener sentido como enunciado.

Distribucionalmente se aprecia una diferencia entre términos y nombres propios por un lado, con valor referencial, y el resto del léxico por otro.

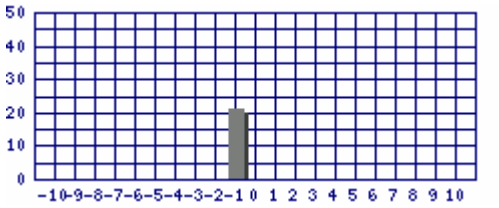
El Trabajo actual

Lo que hemos visto hasta aquí son las relaciones sintagmáticas, las unidades que aparecen en presencia unas de otras. Llamamos a estas coocurrencias de primer grado, frente a las de segundo grado, las que capturan las relaciones paradigmáticas entre las unidades. Las de este segundo tipo no se caracterizan necesariamente por aparecer unas cerca de otras, sino que exhiben un mismo perfil de coocurrencia. Un ejemplo típico es el de los sinónimos, que exhiben constelaciones muy parecidas. Esto hace posible que se puedan extraer automáticamente grupos de cuasisinónimos, como lo hicieron Shuetze y Pedersen (1997).

En nuestro caso vamos a limitar el trabajo al estudio de las relaciones sintagmáticas, aunque de un modo más fino que el hecho hasta ahora, identificando las relaciones de dependencia entre los nodos e identificando los satélites que tiene cada nodo, es decir los vecinos frecuentes en contigüidad o a distancia variable.

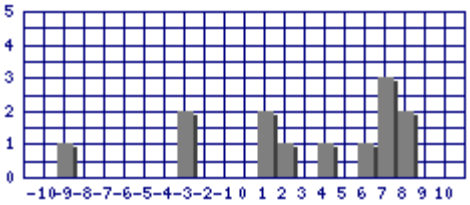
En el mapa conceptual del *inmigrante*, por ejemplo, aparece un nodo *papeles*, pero en realidad la mayor parte de las veces ocurre como *sin papeles*. Hay unas 1180 realizaciones del lema *inmigrante* en el Corpus del IULA. En el subcorpus conformado por estas 1180 oraciones, tres veces aparece la forma *papel* con el sentido de *jugar un papel* y 27 veces aparece la forma plural, *papeles*. Entre estas 27, 21 veces aparece acompañado de *sin*. El siguiente histograma nos muestra en qué posición ocurre *sin* respecto a *papeles*, que es el que está en la posición 0. Las 21 ocurrencias se dan en la posición -1 respecto a papeles.

Forma: **sin**
frecuencia: 21
moda: -1
μ: -1
σ^2: 0
σ : 0



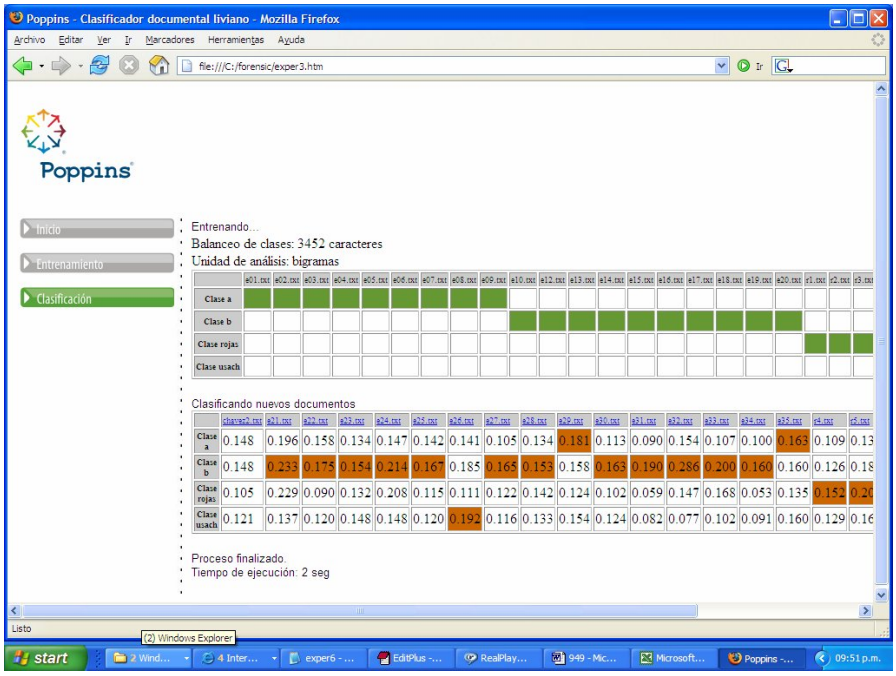
La forma *papeles* no tiene otro compañero tan regular. Las ocurrencias de *en*, por ejemplo, se reparten a izquierda y derecha de *papeles* sin exhibir un patrón regular.

Forma: **en**
Frecuencia: 15
moda: 7
μ: 2.77
σ^2: 16.15
σ: 4.02



Clasificación de documentos

El origen de esta tercera línea de investigación está en un interés que había en el IULA por desarrollar un clasificador de documentos, con la idea de integrarlo a un proyecto que es un sistema de captura de corpus a partir de la web. El problema era que una vez reunida una colección de documentos, para poder ingresarla al corpus era necesario primero clasificar esos documentos según distintos criterios, como la lengua, el tema y el grado de especialización.

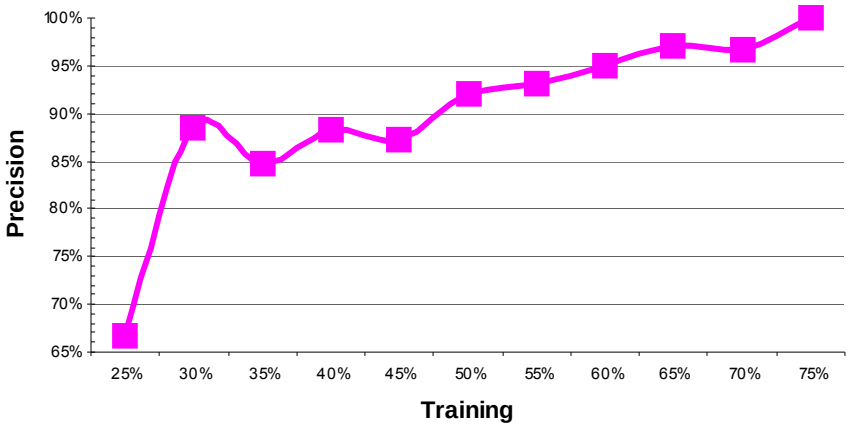


Entonces una primera propuesta fue un algoritmo de aprendizaje supervisado que compara documentos con un *modelo de documento* que el programa ha construido durante una etapa de entrenamiento. La comparación se hace con una medida de similitud basada en los bigramas, entendiendo éstos como secuencias de dos palabras tal como se realizan en el texto, sin ningún tipo de procesamiento lingüístico.

Esta aproximación tiene la ventaja de que el clasificador puede ser de propósito general, aplicable a cualquier criterio, sin que un usuario tenga que especificar cuáles son los atributos que distinguen una categoría de otra. Sólo aprende a través de ejemplos, con una cantidad mínima de 2000 palabras de texto por cada categoría, equivalentes a unas cuatro páginas como esta.

El rendimiento clasificando por tema es de una media de 91% de precisión en experimentos con documentos del CT-IULA.

El mismo experimento se replicó luego clasificando documentos por autor (Nazar y Sánchez Pol, 2007), utilizando el corpus que Sánchez Pol (2006) usó en su proyecto de tesis: unos 100 artículos de opinión, de alrededor de 400 palabras cada uno, escritos en español por cinco autores distintos. Aquí encontramos una atribución correcta de los textos a sus autores en un 92% de las veces, dividiendo la mitad del corpus para entrenamiento y la otra mitad para test. La siguiente es la curva de aprendizaje donde aprecia el aumento de la precisión en la medida que se que eleva el porcentaje del corpus utilizado como entrenamiento.



Otra replicación del experimento fue con los Federalist Papers, famoso caso de autoría disputada, actualmente publicado como ejemplo de funcionamiento en la página web del proyecto[2].

Se trata de una colección de ensayos, escritos en 1788 por distintos autores, que conformaron la base de lo que luego sería la constitución de los Estados Unidos. Doce de ellos tenían una autoría disputada entre Alexander Hamilton y James Madison, pero los estudios estadísticos modernos señalan al segundo como el autor más probable (Mosteller y Wallace, 1984). De la misma manera, nuestro algoritmo clasifica los textos dubitados como escritos por Madison.

El algoritmo

El sistema ordena sus clases de entrenamiento con modelos de enegramas de los textos que conforman cada clase, convirtiendo todos los textos en vectores. Así el vector 1 podría ser en este caso el texto que queremos clasificar y el vector 2 una de las categorías a las que podría corresponder. Pueden haber *n* vectores y *d* dimensiones.

[2] http://www.poppinsweb.com

dimensiones	vector 1	vector 2	etc...
no me	5	0	
, que	4	6	
el hombre	3	2	
. y	3	4	
a su	3	4	
su señora	2	0	
, como	2	3	
me he	2	3	
etc...	...	...	

Ejemplo: fragmento de una matriz de bigramas para comparación de vectores

Antes del aprendizaje, el algoritmo ha hecho una normalización de las clases de entrenamiento llevándolas al tamaño de la clase más chica. Luego la clasificación de nuevos documentos consiste en la conversión de estos documentos en vectores ponderados, en el que los bigramas de este documento aparecen asociados a su frecuencia de aparición. Se compara este vector con los vectores que representan a cada una de las clases y se ubica el nuevo documento en la clase que se le parezca más. La medida de similitud para tal fin es la siguiente:

$$\max_{1 \leq j \leq n} \sum_{i=1}^{d} (x_i^0 + x_i^j)$$

Está emparentada con el Matching Coefficient, pero es aplicable a vectores con valores reales en vez de únicamente vectores binarios, porque selecciona como los más parecidos a aquellos que den el valor más alto cuando se suman los valores de las dimensiones que tienen en común. El Matching (Manning y Shuetze, 1999), en cambio, sólo da un puntaje más alto a los vectores que tienen la mayor cantidad de componentes en común, sin tener en cuenta ningún valor asociado a esos componentes, tal como la frecuencia de aparición.

Evaluación de la significación estadística

La evaluación se hizo con el experimento del corpus de Sánchez Pol (2006), de artículos de opinión. En primer lugar presentamos un test típico, el Pearson's Chi-square test.

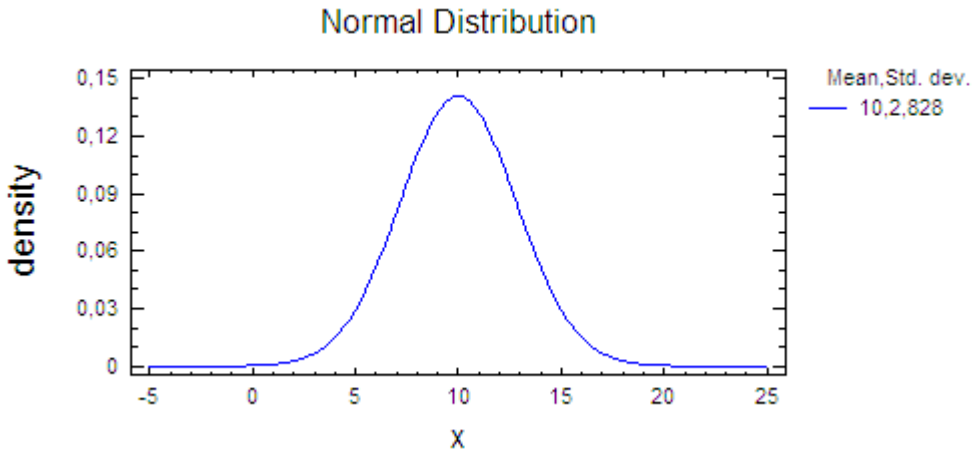
Lo que hacemos con esto es determinar la probabilidad de verdad de una hipótesis nula, que es, este caso, que los documentos son clasificados de manera aleatoria y los resultados son producto de la casualidad. Si esto fuera así, y considerando que tenemos cinco diferentes autores por cada documento a asignar, y que hacemos 50 intentos de clasificación, entonces deberíamos esperar la media de aciertos como una variable aleatoria con una distribución de probabilidad uniforme, con una media aproximada de 10 aciertos. Por el contrario la muestra observada es de 46 aciertos.

La chi-square nos permite comparar la media esperada con los valores realmente observados y determinar qué tan probable es que esta muestra pertenezca a esta distribución uniforme de la hipótesis nula.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Con nuestros números esto da un valor de 162. Con cuatro grados de libertad (número de autores menos uno) esto corresponde a $p < 0.00001$

Otra forma de hacer el test consiste en plotear[3] la distribución de probabilidad que tendríamos si la hipótesis nula fuera cierta.



La gráfica muestra que tendríamos una media esperada de 10 aciertos en una muestra de 50 clasificaciones. Si tuviéramos una muestra de entre 5 y 15 aciertos, o cualquier otro valor por debajo de la curva, este resultado sería perfectamente atribuible al azar. La media de 46 se encuentra fuera de este rango.

[3] Este ploteo se hizo con el software Statgraphics Plus 5.0 (Copyright 1994-200 Statistical Graphics Corp), y con la ayuda del Prof. Jaume Llopis.

Conclusión

El propósito de este seminario ha sido destacar la utilidad de las herramientas de medida que ofrecen los métodos cuantitativos, que pueden aplicarse a estudios totalmente distintos a los comentados aquí.

Lo más interesante, sin embargo, es para mí la aplicación al análisis conceptual. Es porque estamos en un momento único de la historia en que, gracias al desarrollo de tecnología y la transmisión de conocimientos, hoy nos rodea la memoria de la humanidad. Somos infinitamente eruditos, de manera potencial, y actualizamos esa potencia cada vez que hacemos búsquedas en Internet. Paradójicamente, sin embargo, esta disponibilidad de información es tan grande que supera la capacidad de lectura que desarrollamos en la era de Gutenberg, lo que nos impulsa ahora a encontrar nuevas formas de amplificar esa capacidad.

Finalmente, es importante aclarar que en lo que a análisis conceptual se refiere, y a diferencia de los trabajos relacionados con la construcción de ontologías, lo que nos interesa no es lo que las cosas *son*. Lo que nos interesa es lo que la mayoría de las veces se escribe acerca de un determinado concepto, bajo el supuesto de que lo que se dice más frecuentemente acerca de una cosa es un saber establecido dentro de una comunidad discursiva.

Referencias Bibliográficas

Attardi, G., (2006) **Experiments with a Multilanguage Non-Projective Dependency Parser**, Proc. of the Tenth Conference on Natural Language Learning, New York.

Atserias, J., E. Comelles y A. Mayor. (2005) **TXALA un analizador libre de dependencias para el castellano.** Procesamiento del Lenguaje Natural, n. 35, p. 455-456.

Calvo, H. y Gelbukh, A. (2006) **DILUCT: An Open-Source Spanish Dependency Parser based on Rules, Heuristics, and Selectional Preferences.** In: Christian Kop, Günther Fliedl, Heinrich C. Mayr, Elisabeth Métais (eds.). Natural Language Processing and Information Systems. 11th International Conference on Applications of Natural Language to Information Systems, NLDB 2006, Klagenfurt, Austria, May/June 2006.

Church, K., and Hanks, P., (1991), **Word Association Norms, Mutual Information and Lexicography**, Computational Linguistics, Vol 16:1, pp. 22-29, http://research.microsoft.com/users/church/wwwfiles/publications.html

Kilgarriff, A., (1997) **"I don't believe in word senses"**, Computers and the Humanities 31: 91-113.

Mandelbrot, B. (1961). **On the theory of word frequencies and Markovian models of discourse**. Structure of Language and its Mathematical Aspects, Proceedings of the Symposia on Applied Mathematics, v. 12, American Mathematical Society, 190--219.

Manning, C. y Schütze, H. (1999), **Foundations of Statistical Natural Language Processing**. MIT Press.

Marcus, S., Nicolau, E. y Stati, S. (1978). **Introducción a la lingüística**. **matemática**, Barcelona, Ed. Teide.

Mosteller, F y Wallace D. (1984) **Applied Bayesian and classical inference the case of the Federalist papers**, New York, Springer.

Muller, C. (1973), **Estadística Lingüística**, Madrid, Ed. Gredos.

Nazar, R y Sánchez Pol, M. (2007). **An Extremely Simple Authorship Attribution System**, Proc. of the Second European IAFL Conference on Forensic Linguistics, Language and the Law., Barcelona, september 14th- 16th, 2006. (En preparación).

Sánchez Pol, M. (2006), **Proposta d'un mètode d'estilística per a la verificació d'autoria: els límits de l'estil idiolectal**, Projecte de tesi doctoral dirigit per M. Teresa Turell Julià, Barcelona, Universitat Pompeu Fabra.

Schütze, H. y Pedersen, J. (1997). **A cooccurrence-based thesaurus and two applications to information retrieval**. Information Processing and Management. 33(3), p.307-318;.

Shannon, C. (1948) **A Mathematical Theory of Communication,** The Bell System Technical Journal,Vol. 27, pp. 379–423, 623–656, July, October, 1948.

Sinclair, J, (1991) **Corpus, concordance, collocation**, Oxford, Oxford University Press.

Williams, G.C. (1998). **Collocational Networks: Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles**. International Juornal of Corpus Linguistics Vol. 3(1), 151-171, John Benjamins Publishing Co.