La energía textual como medida de distancia en agrupamiento de definiciones

Juan-Manuel Torres-Moreno [1,3], Alejandro Molina [1,2], Gerardo Sierra [2]

[1] Laboratoire Informatique d'Avignon, BP1228, 84911 Avignon Cedex 9, France

[2] Universidad Nacional Autónoma de México, Ciudad universitaria, México D.F.

[3] Ecole Polytechnique de Montréal CP 6079 Succ. Centre Ville H3C 3A7 Montréal (Québec) Canada

Resumen

La consulta de definiciones es una de las tareas mas comunes en los sitios de tipo enciclopédico como Wikipedia, Encarta y Medline. La detección, clasificación y agrupamiento de definiciones son tareas recientemente introducidas y en creciente desarrollo. Estas tareas se complican cuando las definiciones están inmersas en textos recuperados de la Web. Presentamos un algoritmo de clasificación basado en una nueva medida de distancia entre definiciones derivada de la energía textual calculada a partir de una representación vectorial del texto, independiente del idioma. Esta distancia puede tener aplicaciones en agrupamiento de textos cortos como *snippets* y títulos, para los cuales resulta complicado utilizar técnicas clásicas de ponderación como tf-idf porque sus frecuencias son muy bajas. Los resultados obtenidos son bastante alentadores y dan pie a explorar otras propiedades de la distancia propuesta.

Abstract

Definition searching is the most common query in encyclopedic system sites such as Wikipedia, Encarta and Medline. The detection, classification and clustering of definitions are recently introduced tasks in increasing development. These tasks become even more complicated when those definitions are embedded in texts and recovered from the sites as they appear. We present here a clustering algorithm based on a new measure of distance between definitions derived from the textual energy that can be calculated from a text vector representation, which is language independent. The energy distance suggested in this work may also have application for short texts clustering such as snippets and titles, where is difficult to use the classic techniques of weighting as tf-idf since the frequencies of terms are very low. The results are quite encouraging and lead us to explore other properties of the proposed distance measure.

Keywords: clustering, definition extraction, textual energy, neural networks, distance function

1. Introducción

En la actualidad, la cantidad de información escrita es muy abundante debido a la masificación de información en formato electrónico que ha sido propiciada por el uso intensivo de la Web. La proliferación de definiciones en la red, las cuales representan un recurso invaluable para cualquier área del conocimiento, no escapa a esta realidad. Ejemplo de esto es la gran cantidad de acepciones de una misma palabra que pueden encontrarse en Internet. El problema que abordamos en este artículo consiste en agrupar las definiciones de un término de acuerdo con su significado. En la sección 2 hacemos una breve introducción a los contextos definitorios. En 3 presentamos el corpus de evaluación y las técnica de agrupamiento utilizada. En 4 introducimos

el concepto de energía textual y a partir de él deducimos la distancia energética utilizada en un algoritmo de agrupamiento. En la sección 5 evaluamos esta nueva medida como criterio de agrupamiento en un corpus de definiciones en español y su incorporación en un módulo para la presentación de resultados en un sistema de recuperación de información (RI), antes de concluir y presentar algunas perspectivas.

2. Antecedentes

Un contexto definitorio (CD) será todo aquel fragmento textual de un documento especializado donde se define un término (Alarcón et al., 2008). Por ejemplo: *El buque inicialmente lo podemos considerar como un flotador que trata de permanecer en posición vertical frente a perturbaciones exteriores.* El fragmento anterior es un CD debido a que posee el término (T), *buque*, con su respectiva definición (D): *como un flotador que trata* (…). En este ejemplo el término y su definición se relacionan a través de la estructura sintáctica *considerar como*; a este tipo de estructuras que ligan un término con su definición se le llama patrón definitorio (PD). Además de estos elementos existen otros, los cuales ya han sido definidos y establecidos en otros trabajos (Alarcón and Sierra, 2003). En este artículo basta con conocer los elementos mencionados hasta aquí.

El término en un CD es una unidad sobre la cuál se aporta información relevante. Por su parte, la definición en un CD es donde está contenido el conocimiento y la información que se va a transmitir acerca del término con la finalidad de llegar a su comprensión cabal. En este trabajo utilizamos una serie de definiciones provenientes de Internet. Como es de esperar, existe una gran diversidad en el uso del español para transmitir conceptos por medio de definiciones.

Los patrones definitorios pueden ser estructuras sintácticas específicas, como la secuencia *se puede definir como* en: *el índice de aridez se puede definir como el porcentaje de la falta de agua* (Rodríguez, 1999). Cada PD está asociado a un tipo de definición específica. En lo que se refiere al estudio de CDs en español, se plantean 4 tipos de definiciones basadas en el modelo de definición aristotélico: analítica, extensional, funcional y sinonímica (Aguilar, 2009).

Las definiciones analíticas son aquellas que presentan un género próximo que expresa la categoría más general a la cual pertenece el término, así como la información (diferencia específica) que permite distinguir el término de otros elementos de su misma clase. Algunos patrones verbales asociados a las definiciones de este tipo son: *ser+un*, *definir+como*, *entender+como*, *identificar+como*. Las definiciones extensionales enumeran las partes que conforman al término visto como un todo. Algunos verbos ligados a este tipo son: *constar*, *contener*, *comprender*, *incluir*. En las definiciones funcionales no se presenta el género próximo y, en cambio, se introduce una diferencia específica donde se explica la función o el uso particular del término. Algunos patrones relacionados con estas definiciones son: *funcionar*, *encargarse +de*, *permitir*, *servir+para*.

2.1. Extracción automática de CDs

Partiendo de la hipótesis de que los CDs pueden ser extraídos automáticamente mediante la búsqueda de patrones definitorios (Pearson, 1998; Meyer, 2001), Alarcón (2009) presenta el prototipo del sistema *Ecode* (Extractor de Contextos Definitorios): un sistema de reconocimiento y extracción automática de definiciones inmersas en texto libre [1]. Además de la extracción, *Ecode* es capaz de distinguir tres tipos de definiciones de acuerdo con el patrón verbal que éstas presentan. Esta clasificación corresponde con alguno de los tipos: analítica, funcional o

[1] http://brangaene.upf.edu/ecode/.

extensional. Sin embargo, el sistema es incapaz de indicar y distinguir los diversos significados de un término dentro de un mismo tipo de definición. El significado varía según el ámbito en el cual se define el término. Note también que la diversidad de acepciones del término no es distinguible por medio del PD. Por ejemplo: *un virus es un programa*; tiene el mismo patrón definitorio que: *el virus es un microorganismo.*

2.2. Describe: un buscador automatico de definiciones

Describe® (http://www.describe.com.mx) es un sistema de búsqueda de definiciones en español. La extracción de dichas definiciones se realiza a partir de la obtención de fragmentos textuales que incluyan PDs. Después de la extracción, el sistema indiza la información obtenida en la Web, luego la procesa para identificar las definiciones y finalmente muestra al usuario una versión simplificada y debidamente organizada de las definiciones de un término previamente introducido. *Describe* tiene una arquitectura cliente-servidor compuesta por varios módulos que permiten organizar la información disponible en Internet. En este trabajo nos concentramos en el módulo de agrupamiento. Una descripción detallada puede ser consultada en Sierra et al. (2009).

La manera en que *Describe* presenta los resultados está determinada por el módulo de agrupamiento. Su función es estructurar las definiciones de forma que aquellas que poseen significados similares sean asociadas a un mismo grupo y dos grupos distintos hagan referencia a campos semánticos distintos. El proceso de agrupamiento es aplicado para cada tipo de definición aunque al final se muestran en la misma página todos los representantes de cada grupo separados por un encabezado que indica el tipo. Tab. 1 representa una posible (y aceptable) estructura de grupos generada por el módulo de agrupamiento para algunas de las definiciones del tipo analítico del término *virus*. En secciones posteriores se muestran las ideas subyacentes a la implementación del módulo de agrupamiento, haciendo especial énfasis en la medida de distancia entre textos utilizada.

GRUPO 1	*Programa que tiene la capacidad de infectar a otros programas para modificarlos e incluir una copia de si mismo en ellos*
	En tecnología de seguridad en computadoras, un virus es un programa auto replicable que se expande insertando copias de sí mismo en códigos ejecutables o documentos
	Se entiende un programa diseñado para alterar el funcionamiento de los equipos, sin la autorización o conocimiento del usuario
GRUPO 2	*Virus es la máxima expresión de la modernidad en el rock nacional*
	Virus es una banda de rock argentina fundamental del new wave de los años '80, liderada por Federico Moura hasta su muerte, en diciembre del 1988 a causa del VIH. Su hermano Marcelo tomó entonces el rol de vocalista principal y la banda continuó funcionando hasta fines de 1989
GRUPO 3	*El organismo más pequeño que puede causar una infección*
	Microorganismo más pequeño que puede causar infecciones y para sobrevivir necesita estar dentro de una célula viva Tabla 1: Estructura de grupos para las definiciones analíticas del término virus

Tabla 1: Estructura de grupos para las definiciones analíticas del término virus

3. Agrupamiento de definiciones

Agrupar en este trabajo se refiere específicamente a la tarea de utilizar un método de aprendizaje no supervisado para reunir documentos sin incluir información lingüística adicional ni utilizar un conjunto de ejemplos de entrenamiento. Se presentan a continuación las ideas fundamentales de los componentes que integran el algoritmo de agrupamiento semántico de contextos definitorios. Se hace especial énfasis en la definición del criterio de distancia entre las definiciones.

3.1. Obtención de los CDs

La necesidad de crear nuestro propio corpus de experimentación fue inminente. Los CDs son una estructura discursiva recientemente estudiada y los corpus disponibles para su uso son escasos. Además, las características necesarias para la experimentación en este trabajo son muy particulares ya que se centran, por un lado, en la cantidad de acepciones que un término puede tener según su contexto (polisemia) y, por otro, en la probabilidad de extraer suficientes observaciones de la Web. Lo primero resulta evidente si se considera que entre más significados puedan identificarse de un término, más claramente se revelarán los grupos semánticos. Lo segundo se refiere a que es necesario contar con datos suficientes para que los resultados sean representativos. Por estas razones, resultó conveniente seleccionar cuidadosamente los términos a incluir en el Corpus de Términos Polisémicos en Español (CTPE). Con estas ideas, se eligieron primeramente 10 términos que cumplen con las dos características deseadas: son ambiguos o tan generales que su significado está condicionado por el contexto y es muy probable encontrarlos por su uso común. Basándose además en las consideraciones reportadas por Alameda y Cuetos (1995), los términos inicialmente elegidos fueron: *aguja*, *barra*, *cabeza*, *casco*, *célula*, *golpe*, *punto*, *serie*, *tabla* y *ventana*. Después de elegidos los términos surge la cuestión de cómo encontrar CDs en la Web asociados a estos términos. Se decidió combinar la lista de términos elegidos con una lista de patrones definitorios para formar cierto tipo de expresión que denominaremos patrón de búsqueda. Un ejemplo de patrón de búsqueda que combina el término *aguja* y el PD *Ser + Determinante* es: *la aguja es un*. Fig. 1 ilustra algunos patrones de búsqueda con el símbolo <T> representando un término genérico.

la <T> es el	define una <T>
la <T> es un	define a una <T>
...	...
las <T>s son las	definimos una <T>
...	definió una <T>
consideraremos la <T>	...

Figura 1: Patrones de búsqueda

Lo anterior pone de manifiesto la posibilidad de utilizar dichos patrones como expresiones de consulta (*queries*) en motores de búsqueda. La API BOSS de Yahoo! [2] permite extraer referencias de los fragmentos textuales que contienen patrones definitorios en la red. Sin embargo, cabe mencionar que la precisión se vio ceñida por las limitaciones heredadas del sistema de recuperación. Recordemos que un CD es un fragmento de texto donde se define un término mediante el uso de un patrón definitorio. Sin embargo, en la práctica es posible – y muy común – encontrar fragmentos textuales donde aparece un término y solo un contexto de uso que no es, propiamente, una definición. Por ejemplo, en el fragmento: *En general, el miedo a la aguja es el más frecuente.* Observamos que, como en el caso anterior, aparece el término aguja y el patrón *ser+Determinante*, pero evidentemente no se trata de una definición. Decimos que un fragmento de texto es un Candidato a Contexto Definitorio (CCD) si contiene un término y alguna instancia de algún patrón definitorio pero no necesariamente es un CD.

Una vez reunida la información expedida por Yahoo!, decidimos reducir el número de términos en nuestro estudio a 4. La razón principal radica en la gran cantidad de información recuperada

[2] Yahoo! Search BOSS (Build your Own Search Service), http://developer.yahoo.com/search/boss/.

(aprox. 3700 resultados por término en promedio), pues uno de los criterios de evaluación del presente estudio implica la lectura de la información por un humano. Consideramos que basta con conservar algunos de ellos para aplicar el algoritmo de agrupamiento y analizar los resultados. Los términos finalmente seleccionados para conformar el corpus fueron: *barra*, *célula*, *punto* y *ventana*. Tab. 2 muestra la cantidad de CCDs extraídos de la Web para cada término y tipo de definición.

	Analíticas	*Extensionales*	*Funcionales*	*Total*
Barra	1863	307	467	2637
Célula	5352	649	533	6534
Punto	1702	422	750	2874
Ventana	1534	587	565	2686
Total	*10451*	*1965*	*2315*	*14731*

Tabla 2: Candidatos a Contextos Definitorios extraídos de la Web

3.2. Representación binaria de definiciones

Antes de describir la representación, se introducirán algunas definiciones necesarias para la comprensión del resto de la sección. Un documento es una cadena de longitud arbitraria pero finita de símbolos gráficos denominados entidades léxicas (EL). Entendemos como EL aquella que puede ser representada mediante un símbolo o la unión de varios de ellos. Por ejemplo, la palabra *manzana* puede representar una EL, o bien una frase como *Estados Unidos Mexicanos*. Asimismo, una EL puede ser un símbolo ininteligible como *Viv* o *A4*. De esta manera, una colección es un conjunto de documentos y un diccionario es una lista de ELs únicas que aparecen en una colección.

La representación de las definiciones utilizada está basada en el modelo vectorial de textos de Salton (1971). Las definiciones son vectores con tantas entradas como ELs distintas en el diccionario, y el valor de las entradas en cada vector toman el valor de 1 o de 0 dependiendo si la EL correspondiente a una entrada aparece o no en la definición. Así pues, una colección de definiciones se representa por medio de una matriz binaria documento-EL.

3.3. Agrupamiento jerárquico

Un algoritmo de tipo jerárquico ofrece dos grandes ventajas: la primera es que no requiere que el número de grupos sea especificado previamente, la segunda y tal vez la más importante es que tenemos la intuición de que enfatiza relaciones encontradas empíricamente entre las definiciones. Por ejemplo, cuando dos de ellas comparten varias ELs pero una es más específica que la otra. En el algoritmo jerárquico aglomerativo simple (*HAC*), la entrada es un conjunto de objetos y la salida es un dendrograma; es decir, un árbol jerárquico que unifica todos los objetos. En cada iteración, una función calcula la distancia entre cada par de grupos y con esto se determinan los siguientes grupos a combinar para formar un grupo nuevo. El criterio para calcular la distancia entre cada grupo es generalmente una variante de los métodos denominados *linkage*. En el método del vecino mas lejano o *complete linkage* (Sorensen, 1948), la distancia entre dos grupos se representa por la distancia más máxima entre un objeto del primer grupo y un objeto del segundo grupo. Este método tiene la ventaja de generar grupos pequeños, cohesivos y bien delimitados y por ello hemos optado por utilizarlo en nuestro algoritmo. En el método de vecino mas lejano, la distancia entre dos grupos, D_i y D_j, tales que $d_i \in D_i$ y $d_j \in D_j$, se define como en (1):

$$Dist(D_i, D_j) = \underset{d_i \in D_i, d_j \in D_j}{máx} dist(d_i, d_j) \quad (1)$$

Donde *dist* es una función con dominio en el espacio de los objetos (una medida de distancia), a diferencia de *Dist* que tiene dominio en el espacio de los grupos. Independientemente del método *linkage* elegido, el agrupamiento jerárquico converge a la integración de un grupo absoluto. Por tanto, la manera de obtener una partición del conjunto de entrada es realizando algún tipo de corte.

Dos criterios de corte comúnmente utilizados son el corte por umbral de similitud y el corte por distancia. El primero consiste en recuperar como partición del conjunto inicial todos los grupos formados al podar el árbol jerárquico – dendrograma – en alguna altura predeterminada. El segundo considera solamente los grupos cuya distancia de unión se encuentre por debajo de un umbral predeterminado. En la sección siguiente se muestra la manera en que utilizamos un algoritmo jerárquico aglomerativo para reunir una colección de CDs representados en forma vectorial. En dicho algoritmo, la constitución de los grupos se determina por medio de un valor de corte por distancia que denotaremos por α. Nótese que hasta ahora se ha hablado de distancia sin especificar la manera en que ésta se calcula. La similitud o distancia se calcula por medio de una función, la cual toma dos vectores (pertenecientes al mismo espacio) y regresa un valor real en el intervalo [0,1], que cuantifica su semejanza. Existen muchas medidas de distancia, pero nosotros proponemos una medida de distancia nunca antes utilizada, basada en el concepto de energía textual.

4. La energía textual como medida de distancia

El uso de las redes neuronales se ha extendido exitosamente durante los últimos años (Vapnik, 1999; Hastie et al., 2001). La versatilidad de esta herramienta estadística nos lleva desde aplicaciones puramente teóricas (Schölkopf and Smola, 2003) hasta los usos más prácticos imaginables debido a su gran capacidad para reconocer y aprender patrones (Bishop, 1995). El modelo que presentamos en este trabajo parte directamente de uno de los modelos de magnetismo más sencillos de la mecánica estadística (Binder, 2001): el modelo de Ising, que considera espines atómicos dispuestos en una retícula rectangular en el plano X-Y. Inspirado en este modelo, Hopfield (1982) construyó una red neuronal recurrente con capacidad de recuperar patrones a partir de un conjunto de ejemplos y la denominó red de memoria asociativa. En esta red las unidades tienen asociados dos posibles valores de activación (0 ó 1). La configuración de los valores de activación en las unidades determina el estado de la red. Cada estado, a su vez, tiene asociado un número conocido como la energía de la red. Uno de los inconvenientes del modelo de Hopfield es que solamente una fracción de patrones puede ser recordado correctamente, limitando su uso en aplicaciones prácticas.

No obstante, el equipo de investigadores de PLN del LIA [3] observó que el comportamiento de la red de memoria asociativa de Hopfield puede ser ampliamente explotado en aplicaciones de procesamiento de lenguaje (Fernandez et al., 2007a; 2008). Tomando como base el modelo vectorial, se pueden interpretar los elementos de una matriz documento-EL como los espines o neuronas de una red neuronal de Hopfield, al definir los documentos como cadenas de neuronas y haciendo que la neurona *i* esté activa si la EL *i* aparece en la frase o inactiva si está ausente.

En este punto surge una consideración sutil pero importante entre el modelo de Hopfield y la analogía para procesamiento de lenguaje natural: mientras que en las memorias asociativas no

[3] Laboratoire Informatique d'Avignon.

se toma en cuenta la interacción de una unidad consigo misma, en el modelo de energía textual esta interacción es importante. Así, la regla de correlación de Hebb (1949) es muy similar a la versión de Hopfield y se expresa por la ecuación (2); para la cual X es la matriz documento-EL.

$$J = X^T \times X \tag{2}$$

De donde se obtiene que la energía textual de interacción se calcula según (3):

$$E_{textual} = -\frac{1}{2} X \times J \times X^T \tag{3}$$

Que en términos de X se expresa por (4):

$$E_{textual} = -\frac{1}{2} X \times (X^T \times X) \times X^T = -\frac{1}{2} (X \times X^T)^2 \tag{4}$$

Fernandez et al. (2007b) centraron su interés en las relaciones entre términos y frases y denominaron dicha interacción la energía textual de un documento, que ha servido para ponderar las frases en un documento y generar resúmenes automáticos, así como para detectar fronteras temáticas a partir de cambios bruscos en las cadenas de texto, obteniendo en ambos casos buenos resultados (Fernández et al., 2007b; 2008). También han mostrado que la energía textual genera una fuerte correlación según la prueba de Kendall en matrices de grandes cantidades de datos (Fernández et al., 2009). Los resultados también indican que la calidad de los resúmenes es independiente del tamaño de los textos, de los temas abordados y de cierta cantidad de ruido inherente a la lengua. Estas propiedades nos dieron indicios para concebir la energía textual como una función de distancia entre textos.

4.1. La distancia energética

En la ecuación (4) las entradas en la matriz $E_{textual}$ son reales negativos o cero. Puesto que el objetivo es comparar solamente la magnitud de la distancia entre vectores binarios, sin pérdida de generalidad, puede considerarse el valor absoluto de la entradas de la matriz como en (5):

$$E = \left| -E_{textual} \right| \tag{5}$$

En general, puede ocurrir que la máxima energía (la mayor magnitud) no se encuentre en los elementos de la diagonal principal. La interpretación de este hecho, al utilizar la energía como distancia entre textos es que, en ocasiones, habrá frases que son más similares a otras frases que a ellas mismas. Sin embargo, para el corpus de estudio, los elementos máximos sí se concentraron en la diagonal principal (Molina, 2009). Por esta razón no hubo necesidad de conservarlos para futuras comparaciones. Lo anterior permite representar la matriz de energía como un arreglo unidimensional de distancia energética. El arreglo (6) contiene la distancia energética entre cada par de vectores.

$$D_{ener} = \left[e_{12}, e_{13}, e_{14}, ..., e_{1n}, e_{23}, e_{24}, ..., e_{2n}, ..., e_{n-1n} \right] \tag{6}$$

Para restringir los valores de las entradas (normalización) de D_{ener} en el rango [0,1], se utiliza el valor máximo como en la ecuación (7):

$$DistEner_{i,j} = \frac{máx(D_{ener}) - D_{ener}}{máx(D_{ener})} \tag{7}$$

DistEner es un arreglo que contiene la distancia entre cualesquiera dos documentos *i,j* de la colección, dado que E es una matriz simétrica, i.e. $e_{ji}=e_{ij}$. Existe ahora la posibilidad de usar

este arreglo en combinación con un algoritmo de agrupamiento para generar una estructura de grupos utilizando (7) como función de distancia. El módulo de agrupamiento en Describe utiliza un algoritmo de tipo jerárquico aglomerativo y la ecuación (7) para determinar la manera en la que se muestran los resultados de una consulta (Molina, 2009).

5. Resultados

El algoritmo fue ejecutado para todas las colecciones del corpus CTPE: para cada colección que asocia el par (término, tipo de definición). Se creó un programa que varía el parámetro de distancia de corte por distancia desde α=0.1,...,1, con un paso δ=0.01. Sólo son reportados los grupos que reúnen al menos 2 definiciones. Con el fin de que el lector observe todas las definiciones originalmente incluidas en los experimentos, incluimos el grupo cuyo valor de corte es 1. Los resultados completos se pueden consultar en la Web [4].

Para el análisis cuantitativo, la variable independiente es α en todos los casos y las ordenadas son: el número de grupos generados, el recuerdo y la precisión. Como referencia (*baseline*), mostramos las gráficas de las mismas variables pero utilizando la distancia de Hamming (1950) como criterio de distancia entre vectores. La razón por la cual no hemos considerado comparar la distancia con otras medidas como *tf-idf* o el producto coseno es que éstas solo tienen sentido cuando los vectores tienen entradas ponderadas en los números reales. En nuestro caso, las frecuencias de los términos son en su mayoría unitarias, obligándonos a utilizar una medida binaria para la comparación.

5.1. Número de grupos

Al utilizar la distancia energética, el número de grupos aumenta en proporción al valor de corte. El comportamiento de la gráfica es prácticamente el mismo para todos los tipos de definición. Se deduce que el número de grupos generados es independiente del tipo de definición. Fig. 2 ilustra el comportamiento del número de grupos en función del valor de corte por distancia para las definiciones extensionale

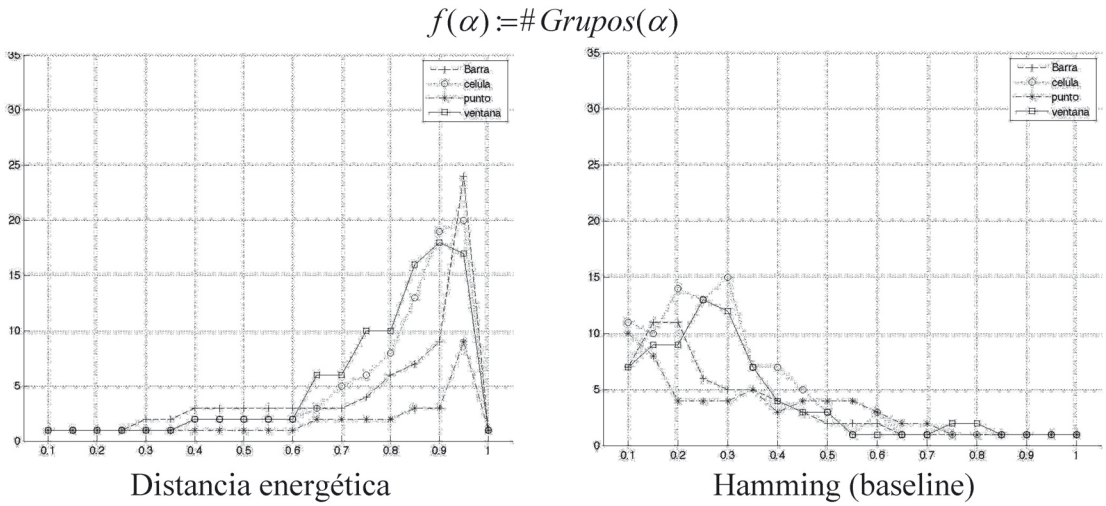


Figura 2: Número de grupos en función del valor de corte α para las definiciones extensionales

[4] http://saussure.iingen.unam.mx/~amolinav/resultados/.

5.2. Precisión y recuerdo

Para los experimentos hemos adaptado dos medidas de evaluación muy populares en recuperación de información: la precisión y el recuerdo (*recall*). La función de recuerdo utilizada fue la proporción de definiciones integradas a algún grupo con respecto al total de las definiciones utilizadas como entrada del algoritmo (ecuación 8). Es decir, cuántas definiciones logramos integrar en el agrupamiento. El recuerdo devuelve un valor entre 0 (si no se formó ningún grupo con al menos 2 definiciones) y 1 (si todas las definiciones fueron integradas en algún grupo).

$$r = \frac{\left|\{total\_definiciones\} \cap \{definiciones\_asignadas\_a\_un\_grupo\}\right|}{\left|\{total\_definiciones\}\right|} \tag{8}$$

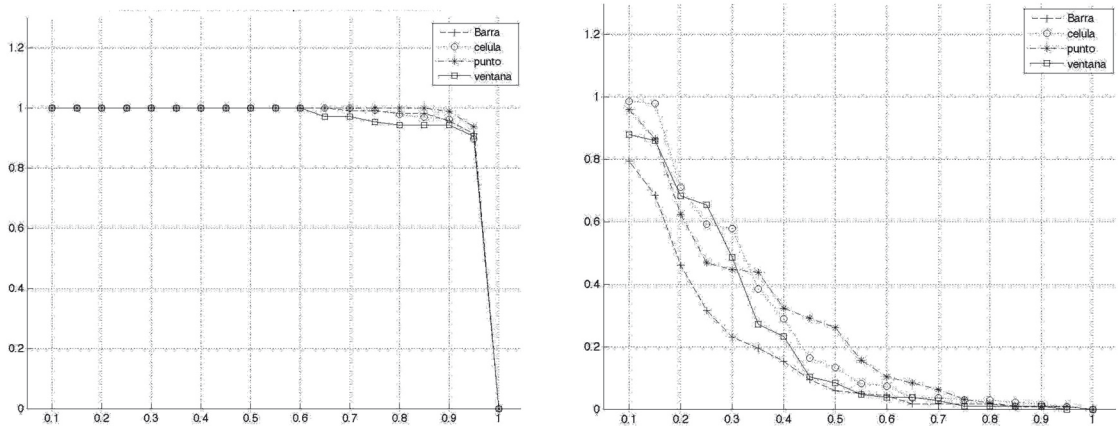
La precisión, por su parte, se define como la proporción de intrusos en un agrupamiento generado (ecuación 9). Ella indica cuántos errores se cometen al integrar los grupos. La precisión devuelve un valor entre 0 (si no es posible determinar la acepción de ninguno de los grupos) y 1 (si ningún grupo contiene intrusos).

$$p = \frac{\left|\{definiciones\_asignadas\_a\_un\_grupo\} - \{Intrusos\}\right|}{\left|\{definiciones\_asignadas\_a\_un\_grupo\}\right|} \tag{9}$$

La precisión se evaluó al comparar los resultados del algoritmo contra una clasificación manual previa. Cabe mencionar que una sola persona realizó la evaluación y que el *golden standard* utilizado fue exactamente el mismo para la distancia energética y para el *baseline*.

Dado que los resultados son similares para todos los tipos de definición; mostramos solamente los resultados de las definiciones extensionales. Fig. 3 ilustra el comportamiento de la precisión en función del valor de corte por distancia para las definiciones extensionales y figura 4 ilustra el comportamiento del recuerdo.

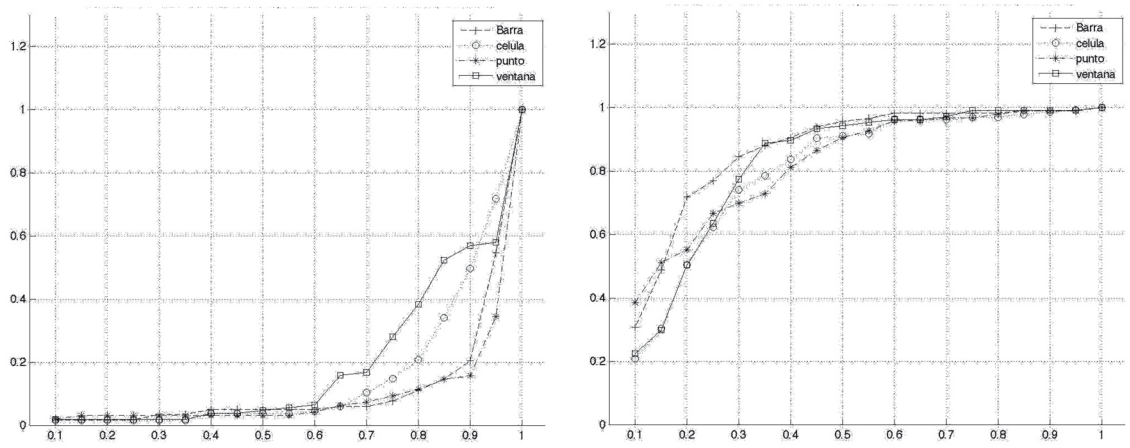
$$f(\alpha) \coloneqq precisión(\alpha)$$



Distancia energética Hamming (baseline)

Figura 3: Precisión en función del valor de corte α para las definiciones extensionales

$$f(\alpha) := Recuerdo(\alpha)$$



Distancia Energética Hamming (baseline)

Figura 4: Recuerdo en función del valor de corte α para las definiciones extensionales

5.3. Análisis cualitativo

Los resultados a nivel cualitativo son alentadores. Se generan agrupamientos que reflejan las distintas acepciones que tiene un término, pero también se distinguen sutilezas existentes dentro de una misma acepción. Considere como ejemplo de este fenómeno las definiciones 1 y 2 agrupadas para el término *célula*, de tipo extensional, con valor de corte α=0.8:

Definición 1. La célula se compone de un núcleo envuelto en protoplasma, alrededor del cual hay una membrana que separa la célula de su medio ambiente.

Definición 2. La célula consta de una membrana celular que envuelve una masa viscosa y granulosa llamada protoplasma, en la cual se encuentran todos los organelos celulares, incluido el núcleo.

En la primera definición se infiere que la célula se compone de un núcleo que está envuelto en protoplasma y que a su vez el protoplasma está envuelto por una membrana. La segunda es estructuralmente inversa pero semánticamente equivalente. Esto es, la célula consta de una membrana que envuelve al protoplasma en el que está el núcleo.

Otro resultado importante es que, a medida que α tiende a 1, los grupos se vuelven más específicos y a veces indeseablemente explícitos en el significado de las palabras que los conforman.

6. Conclusiones y trabajo futuro

Se puede considerar que, independientemente del tipo de definición, el comportamiento del algoritmo puede ser dividido en tres zonas según el valor de corte: zona 1 para $0 \leq \alpha \leq 0.7$. En esta zona se obtiene muy alta precisión (>90%) y bajo recuerdo (<40%). Se debe utilizar un α en este intervalo cuando se quieran obtener acepciones comunes, pocos grupos (~5) con pocas definiciones y sin la presencia de intrusos en los grupos generados. La zona 2, con $0.75 \leq \alpha \leq 0.85$ se caracteriza por tener precisión alta (~80%) y recuerdo intermedio (~50%). Este intervalo es de mayor interés dado que representa un buen balance entre la precisión y el número de grupos generados (~10). Por último, la zona 3, para $0.85 \leq \alpha \leq 0.99$, proporciona precisión media (~50%) y recuerdo alto (~80%). El número de grupos generados en esta zona

es alto (~20) pero cada grupo es muy preciso en el significado. Por tanto, el valor de corte por distancia recomendado para el algoritmo debe ser $0.75 \leq \alpha \leq 0.85$. Una posible mejora del módulo de agrupamiento sería determinar de manera dinámica el α que genera la mejor estructura de grupos.

Por último, cabe mencionar que la fórmula de energía textual utilizada pondera las relaciones entre frases con caminos de orden 2, es decir, relaciona frases que tienen términos en común y aquellas que tienen vecinos en común, aún cuando éstas no compartan ningún término. No obstante, es posible generalizar el concepto y obtener las matrices de adyacencias de potencias mayores a 2. En general $E=(X \times X^T)^N$ representa las interacciones de orden N entre las frases representadas en una matriz X. En este trabajo nos hemos concentrado exclusivamente en el análisis para el orden 2, dejando la posibilidad de experimentación con ordenes superiores a trabajos futuros.

Agradecimientos

Este artículo fue parcialmente financiado por el proyecto de *Extracción de relaciones léxicas para dominios restringidos a partir de contextos definitorios en español* (registro: 82050, CONACYT-México 2009).

Referencias

Aguilar C. (2009). *Análisis lingüístico de definiciones en contextos definitorios*. Tesis de doctorado, México: UNAM.

Alameda J. and Cuetos F. (1995). *Diccionario de Frecuencia de las unidades lingüísticas del castellano*. Servicio de Publicaciones, Universidad de Oviedo.

Alarcón R. (2009). *Extracción automática de contextos definitorios en corpus especializados. Propuesta para el desarrollo de un ECODE*. Tesis de Doctorado, Barcelona: IULA, Universidad Pompeu Fabra.

Alarcón R., Bach C. and Sierra G. (2008). Extracción de contextos definitorios en corpus especializados: Hacia la elaboración de una herramienta de ayuda terminográfica. *SEL*, Vol. (37): 247-278.

Alarcón R. and Sierra G. (2003). El rol de las predicaciones verbales en la extracción automática de conceptos. *Estudios de Lingüística Aplicada*, Vol. (38): 129-144.

Binder K. (2001). *Encyclopaedia of Mathematics*. Dordrecht: Kluwer Academic Publishers.

Bishop C.M. (1995). *Neural networks for pattern recognition*. Oxford: Oxford University Press.

Fernández S., SanJuan E. and Torres-Moreno J.-M. (2007a). Textual Energy of Associative Memories: performants applications of ENERTEX algorithm in Text summarization and topic segmentation. In *Proceedings of MICAI 2007*, pp. 861-871.

Fernández S., SanJuan E. and Torres-Moreno J.-M. (2007b). Energie textuelle de mémoires associatives. In *Conference TALN 2007*, pp. 25-34.

Fernández S., SanJuan E. and Torres-Moreno J-M. (2008). Enertex: un système basé sur l'énergie textuelle. In *Conference TALN 2008*.

Fernández S., SanJuan E. and Torres-Moreno J.-M. (2009). Résumés de textes par extraction de phrases, algorithmes de graphe et énergie textuelle. In *XVI-èmes rencontres de la Société Francophone de Classification*, Grenoble, pp. 101-104.

Hamming R. (1950). Error detecting and error correcting codes. *Bell System Technical Journal*, Vol. (2): 147-160.

Hastie T., Tibshirani R. and Friedman J. (2001). *The elements of statistical learning*. New York: Springer-Verlag.

Hebb D. (1949). *The organization of behavior*. New York: Wiley.

Hernández A. (2009). *Análisis lingüístico de definiciones analíticas para la búsqueda de reglas que permitan su delimitación automática*. Tésis de licenciatura, México: UNAM.

Hopfield J. (1982). Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the National Academy of Sciences of the USA*, Vol. (9), pp. 2554-2558.

Meyer I. (2001). Extracting Knowledge-rich contexts for Terminography. In Bourigault, D. and Jacquemin, C., editors, *Recent advances in Computational Terminology*, Amsterdam: John Benjamins, pp. 279-302.

Molina A. (2009). *Agrupamiento semántico de contextos definitorios*. Tesis de maestría, México: UNAM.

Pearson J. (1998). *Terms in context*. Amsterdam: John Benjamins.

Rodríguez C. (1999). Operaciones Metalingüísticas Explícitas en Textos de especialidad. Trabajo de investigación, Barcelona: IULA, Universitat Pompeu Fabra.

Salton G. (1971). *The SMART Retrieval System: Experiments in automatic document procesing*. Englewood Cliffs: Prentice Hall.

Schölkopf B. and Smola A.J. (2003). *Learning with kernels*. Cambridge (MA): The MIT Press.

Sierra G., Alarcón R., Molina A. and Aldana E. (2009). Web Exploitation for Definition Extraction. In *Proceedings of LA-WEB 2009* (7th Latin American Web Congress).

Sorensen T. (1948). A method of estimating groups of equal amplitude. *Plant sociology based on similarity of species content*, Vol.(5): 1-34.

Vapnik V.N. (1999). *Statistical learning theory*. New York: Wiley.