Determinación de relaciones léxicas con base en el grado de subsunción*

Juan Fajardo González & Héctor Jiménez Salazar
Facultad de Ciencias de la Computación
B. Universidad Autónoma de Puebla
C.U. 72570, Puebla, México
tel. 01(222)2295500 ext. 7212
hjimenez@fcfm.buap.mx

Resumen

Este trabajo presenta los resultados de la aplicación de una medida de subsunción entre un par de acepciones de palabras del español, usando un *corpus* sin mayor preprocesamiento que la extracción de seudolexemas de las palabras de clases abiertas. Con base en las propiedades de los conceptos formales, la subsunción, vista como una relación entre palabras a partir del uso de éstas en una colección de textos, es reexpresada en términos de las características de los textos. Ya que la subsunción se apoya en la contención (como operación de conjuntos) de una colección de textos en otra, ésta se aborda en forma aproximada. Se trabaja, entonces, con el grado de subsunción. Algunas relaciones léxicas, como la sinonimia y la hiponimia pueden derivarse a partir de los grados de subsunción calculados entre dos palabras.

*El presente es resultado de pruebas adicionales realizadas con base en el trabajo presentado durante CICLing 2003.

Palabras clave:

Relación semántica, subsunción, conceptos formales.

Abstract

This work presents the results of applying of a measure to deal with word senses obtained from a raw *corpus*. In order to determine some lexical relationship between two word senses formal concept theory and the notion of subsumption are used to define the subsumption ratio. Here we use a set of texts to represent the use of a word sense. The main idea is to quantify how much a set is contained in another to conclude a kind of lexical relationship.

Keywords:

Semantic relationship, subsumption, formal concepts.

1. Introducción

Los Sistemas de Recuperación de Información (SRI) usan las relaciones léxicas para mejorar su desempeño. En particular, los SRI aplican técnicas de *expansión de consultas* para efectuar un enriquecimiento de las peticiones de los usuarios (Mandala *et.al.*, 1999). Es por ello importante abordar el problema de construir bases de datos léxicas de diferentes dominios (Grefenstette, 1993). En (Sanderson & Croft, 1999), se define que la relación x es hipónimo de y apoyándose en la noción de subsunción: x *subsume* y si $A_y \subset A_x$, donde A_w denota el conjunto de contextos (documentos, oraciones, etc.) que contienen la

palabra w. Sin embargo, se sabe que, en general, para hipónimos $A_y \subset A_x$ no se satisface estrictamente: solamente 80 % de los miembros de A_y están en A_x (Sanderson & Croft, 1999).

Es de nuestro interés la exploración de diversas relaciones léxicas. En este trabajo enfrentamos el problema de contención de un conjunto en otro, derivado de la noción de subsunción, mediante el cálculo de la proporción de contención. A diferencia del uso de contextos para verificar la subsunción, usamos las características de la acepción de una palabra; esto es, las palabras que ocurran en el contexto, tomado éste como la oración en que ocurra la palabra. Las características de una entidad se tratan dentro del marco de los conceptos formales (Davey & Pristley, 1990) haciendo uso de la idea de índice definido para un SRI.

En la teoría de conceptos formales, un concepto está definido por una pareja formada por el *extento*, el conjunto de ejemplares que presentan el concepto, y el *intento*, el conjunto de características que satisfacen todos los ejemplares del concepto. Asimismo, se dice que un concepto, con extento A_x e intento B_x, (A_x, B_x), es *más particular* que (A_y, B_y), expresado como $(A_x, B_x) \leq (A_y, B_y)$, si y sólo si $A_x \subset A_y$, o equivalentemente $B_y \subset B_x$ (Davey & Pristley, 1990).

El enfoque, entonces, frente a la subsunción es tratar como concepto formal a cada posible acepción de una palabra; el extento estará formado por el conjunto de contextos donde se use la palabra en una acepción, y el intento corresponderá a las palabras "más representativas" del extento, según la noción de *valor discriminante* en los SRI.

En la siguiente sección se precisa el término *grado de subsunción*, enseguida un experimento llevado a cabo para conocer el funcionamiento de esta medida y, al final, las

conclusiones de este trabajo.

2. Grado de subsunción

Consideremos una colección de textos $\mathcal{T} = \{T_1, T_2, \ldots, T_N\}$, y una palabra x. Si la palabra x está contenida en el texto T_i, éste lo podemos ver como un ejemplar del concepto "el uso de la j-ésima acepción de la palabra x", para alguna j; concepto que denotaremos en lo sucesivo como x_j. Sea $\mathcal{T}_x = \{T_{x1}, T_{x2}, \ldots T_{xm}\} \subset \mathcal{T}$ la colección de textos que usan la palabra x. Podemos representar cada T_{xi} por sus términos índice, esto es, las palabras más representativas de T_{xi} con respecto a $\mathcal{T}_x$. Usamos para este propósito el modelo de valor discriminante (Salton *et.al.*, 1975): dada la frecuencia inter-documento de un término v (o el número de documentos que emplean v), denotado por df_v, los términos representativos y contenidos en T_{xi} respecto a $\mathcal{T}_x$ son aquellos que satisfacen que $\frac{m}{100} \leq df_v \leq \frac{m}{10}$. Lo anterior es la base para determinar el intento de un ejemplar. Para obtener el intento de un concepto x_j, antes debemos identificar los ejemplares del extento de x_j. Así, hemos de agrupar los elementos de $\mathcal{T}_x$ según su similitud, bajo la hipótesis de que los elementos similares usan x en una acepción similar. Agrupamos entonces la colección $\mathcal{T}_x$ con la función de similitud de Jaccard:

$$sim(T_{xi}, T_{xj}) = \frac{\#(T_{xi} \cap T_{xj})}{\#(T_{xi} \cup T_{xj})}, \tag{1}$$

donde $\#X$ representa la cantidad de elementos del conjunto X. Fueron probados dos criterios para agrupar la colección de textos:

1. Considerando aquellas instancias que son más similares para formar un grupo, y

2. Tomando aquellos grupos que son más numerosos.

El primer criterio obtuvo mejores resultados (Jiménez, H., 2003). El procedimiento para agrupar la colección de textos tomó, entonces, como criterio: formar un grupo con aquellos textos que son igualmente o más similares que el promedio total de las similitudes. Tenemos, así, que $\mathcal{T}_x = \mathcal{T}_{x_1} \cup \mathcal{T}_{x_2} \cup \ldots \cup \mathcal{T}_{x_p}$, donde cada $\mathcal{T}_{x_j}$ es el extento del concepto x_j. Dado un extento es posible construir su intento siguiendo la definición. El intento del concepto x_j, $\mathcal{T}'_{x_j}$, es el conjunto de las palabras de frecuencia intermedia (términos con mayor valor discriminante) usadas por todos los textos contenidos en $\mathcal{T}_{x_j}$.

A partir de lo anteriormente expuesto, podemos obtener el intento de la acepción de dos palabras, $\mathcal{T}'_{x_j}$, $\mathcal{T}'_{y_k}$, y comprobar si $\mathcal{T}'_{y_k} \subset \mathcal{T}'_{x_j}$ para concluir que la j-ésima acepción de la palabra x es un hipónimo de la k-ésima acepción de la palabra y. Ya que, al comparar conjuntos de textos, es frecuente la contención parcial de un conjunto en otro, definimos el grado de contención con la fórmula:

$$\rho(y_k, x_j) = \frac{\#(\mathcal{T}'_{y_k} \cap \mathcal{T}'_{x_j})}{\#\mathcal{T}'_{y_k}}. \tag{2}$$

Un valor alto de $\rho(y_k, x_j)$ significa que una proporción alta de las características de la acepción y_k son parte de las características de la acepción x_j. Además, si $\rho(x_j, y_k)$ tiene un valor bajo, tendremos que sólo una pequeña parte de las características de la acepción x_j se comparten con las de la acepción y_k. Con las dos condiciones anteriores cubiertas, podemos decir que aproximadamente x_j tiene más características que y_k, o bien que la acepción x_j es más restrictivo que la acepción y_k. En suma, un valor alto de $\rho(y_k, x_j)$ y un valor bajo

de $\rho(x_j, y_k)$ indica que y_k subsume a x_j y, por tanto, la palabra x en su j-ésima acepción es un hipónimo de la palabra y en su k-ésima acepción. Las diferentes combinaciones de tipos de valores de $\rho(y_k, x_j)$ y $\rho(x_j, y_k)$ pueden determinar algunas relaciones semánticas, como las que muestra la tabla 1. Esta serie de reglas maneja implícitamente dos umbrales, μ_1, el máximo para los valores "bajos" de ρ, y μ_2, el máximo para los valores "medios" de ρ.

Relación	$\rho(y_k, x_j)$	$\rho(x_j, y_k)$
x_j sinónimo de y_k	alto	alto
x_j hipónimo de y_k	alto	bajo
x_j en relación fuerte con y_k	alto	medio
x_j en relación débil con y_k	bajo	medio
x_j sin relación con y_k	bajo	bajo

Cuadro 1: Condiciones para determinar la relación léxica entre palabras.

3. Determinación de algunas relaciones léxicas

El *corpus*[1] es un conjunto de 2057 textos, con un total de 61,216 oraciones, un vocabulario de 136,988 signos (palabras diferentes incluyendo puntuación, abreviaciones y números), en total 18,092 seudolexemas (obtenidos por el empleo de un truncador de Porter adaptado al español sin ser aplicado a los nombres propios) y alrededor de diez millones de caracteres. De esta colección de documentos fueron retiradas las palabras cerradas y se aplicó a las restantes un algoritmo de truncamiento para dejar sólo los seudolexemas. Se eligieron las siguientes palabras (con su frecuencia entre paréntesis) para efectuar una prueba de clasificación usando ρ: `triunfo` (120), `victoria` (140), `militar` (520), `coronel` (251),

[1] *80 Años Informando (1916-1996)*, colección de artículos del periódico *El Universal*.

`teniente` (100), `avión` (590), `aeroplano` (62), `aeropuerto` (166) e `hijo` (507). Esta selección fue resultado de tomar aquellas palabras con frecuencia suficiente para proporcionar contextos y que cuya relación entre los pares que se formaran fuera evidente (por ejemplo `aeroplano` y `avión`). Se calculó para cada palabra x de la lista anterior: $\mathcal{T}_x$, la colección de textos que usan a x, $\{\mathcal{T}_{x_1}, \ldots, \mathcal{T}_{x_p}\}$; la partición de $\mathcal{T}_x$ de acuerdo con el uso de una acepción de x, y para cada grupo de la partición se calculó su intento: $\{\mathcal{T}'_{x_1}, \ldots, \mathcal{T}'_{x_p}\}$. Dadas las particiones de dos palabras, x de tamaño p, e y de tamaño q, se formaron todos los pares posibles entre los elementos de las particiones y se les aplicó ρ: $\rho(\mathcal{T}'_{x_1}, \mathcal{T}'_{y_1}), \ldots \rho(\mathcal{T}'_{x_p}, \mathcal{T}'_{y_q})$. Asimismo, se realizó el cálculo simétrico $\rho(\mathcal{T}'_{y_1}, \mathcal{T}'_{x_1}), \ldots \rho(\mathcal{T}'_{y_q}, \mathcal{T}'_{x_p})$. Con estos valores pudo apreciarse cuál es la relación dominante entre el par de palabras. La tabla 2 muestra la diferencia máxima entre los pares simétricos de ρ, Δ_{max}, lo cual indica que el par de acepciones considerado es representativo de la clase de relación semántica entre las palabras. También esta tabla contiene el número de cada grupo (acepción) correspondientes a x e y que fueron combinados (columna 3); el cálculo del grado de subsunción (columnas 4 y 5), y la clase obtenida para el par de palabras.

x-y	Δ_{max}	j,k	$\rho(x_j, y_k)$	$\rho(y_k, x_j)$	Clase
`triunfo-victoria`	0.09	1,1	0.54	0.45	sinonimia
`militar-teniente`	0.40	4,2	0.50	0.08	hiperonimia
`militar-coronel`	0.48	3,2	0.66	0.18	hiperonimia
`aeroplano-avión`	0.63	2,1	0.02	0.66	hiponimia
`aeropuerto-avión`	0.54	2,2	0.31	0.85	rel. fuerte
`aeropuerto-aeroplano`	0.31	2,2	0.36	0.05	rel. débil
`hijo-aeropuerto`	0.07	4,2	0.18	0.10	sin relación

Cuadro 2: Clasificación con el grado de subsunción.

En la tabla, hemos manejado los umbrales $\mu_1 = 0{,}2$ and $\mu_2 = 0{,}4$. En las figuras 1, 2

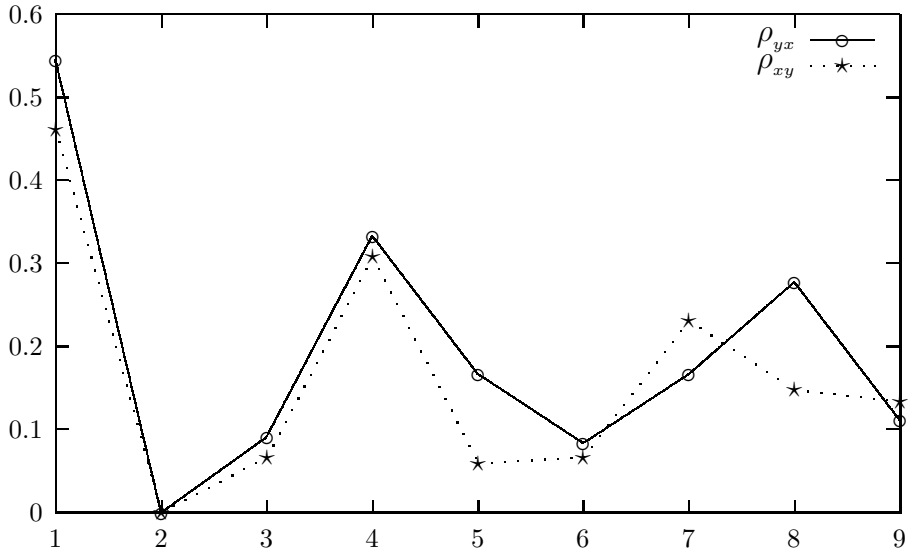

Figura 1: x =`triunfo` y y =`victoria`

y 3 se presentan las gráficas de $\rho(x_j, y_k)$ y $\rho(y_k, x_j)$ para las palabras indicadas. En cada una el número de combinación entre dos acepciones aparece en el eje horizontal, mientras que en el eje vertical se consideran los valores de ρ. Por ejemplo, en la gráfica de la figura 2, el número de grupos más grandes para x =`militar` fue de seis, y para y =`teniente` fue dos. Se obtienen así doce combinaciones posibles entre las acepciones de estas palabras. Cada figura muestra los grados de contención entre los intentos que representan a las diferentes acepciones y qué par de acepciones, si existe, es el más determinante.

4. Conclusiones

En este trabajo hemos definido una medida basada en la noción de subsunción y la teoría de conceptos formales para conocer la relación entre las acepciones de dos palabras. La ventaja de este enfoque es que no requiere más que un *corpus* sin información adicional.

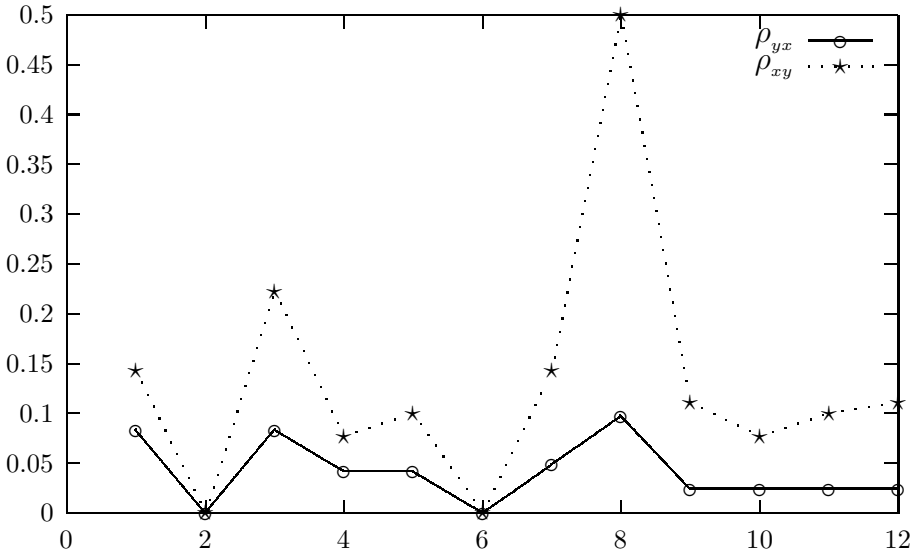

Figura 2: x =`teniente` y y =`militar`

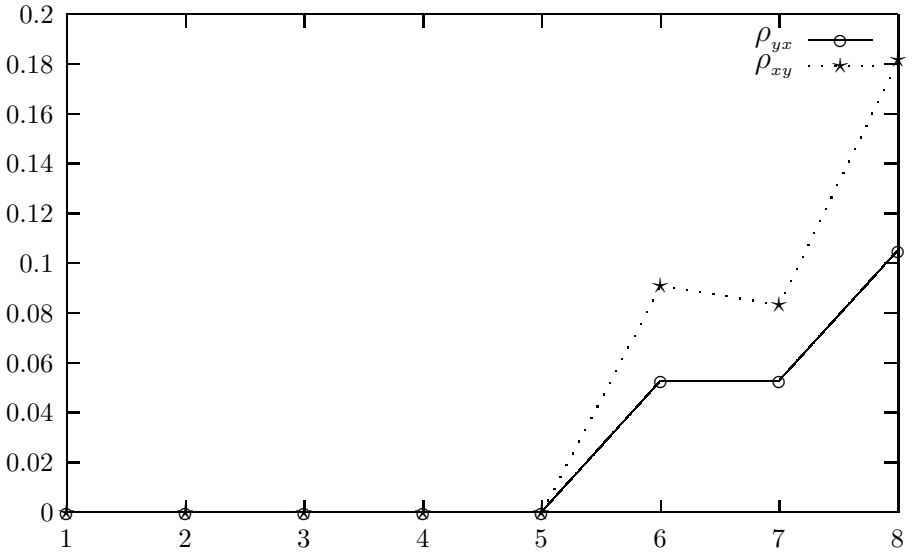

Figura 3: x =`aeropuerto` y y =`hijo`

Sin embargo, es necesario realizar pruebas exhaustivas con el fin de dar mayor sustento al empleo de lo que aquí se ha propuesto para determinar relaciones léxicas.

Agradecimientos

Los autores desean agradecer al Consejo Nacional de Ciencia y Tecnología por apoyar este trabajo a través del proyecto I39165A. Asimismo, a los árbitros por sus valiosos comentarios, y a la Dra. Elena Carcedo por su esmerada revisión.

Referencias

DAVEY, B. & PRIESTLEY, H. (1990) *Introduction to lattices and order*. Cambridge Mathematical Textbooks.

GREFENSTETTE, G. (1993) "Automatic thesaurus generation from raw text using knowledge-poor techniques". En *Making sense of words, 9th. Annual Conference of the UW Centre of the new OED and text Research.*

JIMÉNEZ SALAZAR, H. (2003) "A Method of Automatic Detection of Lexical Relationships using a Raw Corpus", en *Lecture Notes in Computer Science*, Vol. 2588, (Ed.) A. Gelbukh, Springer, 325-328.

MANDALA, R.; TOKUNAGA, T. & TANAKA, H. (1999) "Combining multiple evidence from different types of thesaurus". En *Proc. 22nd International Conference ACM-SIGIR*, 191-197.

SALTON, G.; YANG, C.S. & YU, C.T. (1975) "A theory of term importance in au-

tomatic text analysis", En *Journal of American Society for Information Science*, 26(1), 33-44.

SANDERSON, M. & CROFT, W.B. (1999) "Deriving concept hierarchies from text", En *Proc. 22nd International Conference ACM-SIGIR*, 206-213.