Procesamiento del Lenguaje Natural, Revista nº 64, marzo de 2020, pp. 29-36 recibido 08-11-2019 revisado 15-01-2020 aceptado 18-01-2020

Tarántula –> araña –> animal: asignación de hiperónimos de segundo nivel basada en métodos de similitud distribucional

Tarantula –> spider –> animal: second level hypernymy discovery based on distributional similarity methods

Rogelio Nazar, Javier Obreque, Irene Renau
Instituto de Literatura y Ciencias del Lenguaje
Pontifica Universidad Católica de Valparaíso
rogelio.nazar@pucv.cl, j.obrequezamora@gmail.com, irene.renau@pucv.cl

Resumen: La asignación automática de hiperónimos sigue presentando problemas para el procesamiento del lenguaje natural. En particular, los sustantivos polisémicos se vinculan a distintos hiperónimos y por ello pueden causar problemas estructurales en una taxonomía léxica. Por ejemplo, el sustantivo *tarántula* puede ser registrado como hipónimo de *araña* y, como este es un sustantivo polisémico (puede denotar a un ser vivo o a un tipo de lámpara), es necesario determinar cuál es el hiperónimo siguiente en la cadena: *animal* o *artefacto*. En el presente artículo exploramos métodos para resolver este problema utilizando el cálculo de la similitud entre sustantivos utilizando como variable predictora los verbos con los que coocurren. Los mejores resultados (84 % de acierto) se obtienen con un método simple que solo mide coocurrencia, sin tener en cuenta información sintáctica.
Palabras clave: hiperonimia, polisemia, similitud distribucional, taxonomía

Abstract: Automatic hypernymy discovery continues to present challenges for natural language processing. Polysemous nouns are linked to more than one hypernym and can therefore cause structural damage on a lexical taxonomy. For instance, the Spanish noun *tarántula* ('tarantula') is a hyponym of *araña* ('spider'), but this is also a polysemous noun, as it means 'chandelier' as well. It is thus necessary to determine the next hypernym in the chain, that is *animal* ('animal') or *artefacto* ('artifact'). In this paper we explore methods to solve this problem using a similarity measure that uses verb-noun co-occurrence as a predictor variable. Best results (84 % success) are obtained with a simple method that only measures co-occurrence, irrespective of any syntactic information.
Keywords: distributional similarity, hypernymy, polysemy, taxonomy

1 Introducción

El establecimiento de relaciones de hiperonimia entre unidades léxicas continúa siendo un desafío en el campo del procesamiento del lenguaje natural. Actualmente, las estrategias que existen alcanzan promedio de precisión que fluctúa en torno al 80 % (Velardi, Faralli, y Navigli, 2013; Bordea, Lefever, y Buitelaar, 2016), lo cual deja un amplio margen de mejora.

En el marco de la inducción automática de taxonomías léxicas, es decir, las estructuras que emergen de las relaciones de hiperonimia-hiponimia (Lyons, 1977), uno de los problemas pendientes es cómo tratar adecuadamente el fenómeno de la polisemia (Bordea et al., 2015; Klapaftis y Manandhar, 2010). Eso es así, al menos, en el caso de las taxonomías semasiológicas, es decir, aquellas que se basan en unidades léxicas, como es usual en lexicografía, y no en conceptos, que es lo propio de las ontologías (Baldinger, 1977; Sager, 1990).

La polisemia es el fenómeno por el cual una palabra tiene más de un significado, y ocurre cuando uno de los significados da origen a otro u otros, por metáfora, metonimia u otro mecanismo, sin que el significado original se anule (Ullmann, 1972; Lyons, 1977; Kilgarriff, 1992; De Miguel, 2016). Así, de *araña* 'animal' se deriva *araña* 'lámpara' por la similitud formal entre ambas entidades.

En este trabajo se aborda específicamente el problema de la herencia semántica en la

ISSN 1135-5948. DOI 10.26342/2020-64-3

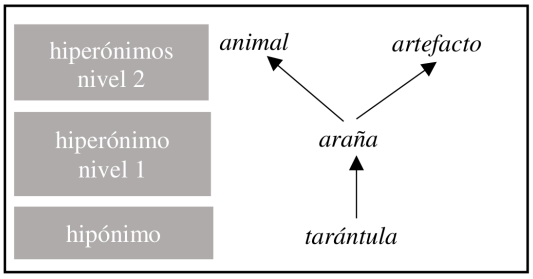

Figura 1: Estructura taxonómica con un hiperónimo de nivel 1 y 2 de nivel 2.

cadena de relaciones de hiperónimo-hipónimo en una taxonomía léxica. Con el fin de ilustrar la problemática, imaginemos el caso de un algoritmo de asignación de hiperónimos que establece correctamente la relación de hiponimia del sustantivo *tarántula* con respecto a *araña*. Tal como se muestra en la Figura 1, sería necesario entonces determinar de qué significado específico del sustantivo *araña* se trata, ya sea el de *animal* (solución correcta) o el de *artefacto* (solución incorrecta). En este trabajo, se llamará *hiperónimo de nivel* 1 al hiperónimo del tipo *araña* (inmediatamente superior en la cadena hiperonímica al hipónimo, en este caso *tarántula*), e hiperónimo de nivel 2 al hiperónimo del tipo *animal* o *artefacto* (de los cuales solo uno de ellos es correcto, en este caso *animal*).

El objetivo de esta investigación es, entonces, proponer un algoritmo para resolver los casos de ambigüedad de hiperónimos de nivel 2. El método está basado en medidas de coocurrencia léxica, a través de las cuales es posible seleccionar el significado correcto entre los que ofrece un hiperónimo polisémico. Para ello, en esta investigación empleamos la coocurrencia sustantivo-verbo. Así, los verbos con los que frecuentemente coocurre el sustantivo *tarántula* proporcionarán pistas sobre si se debe clasificar como *animal* o como *artefacto*. Consideramos que la conformación de este método representa un avance en el marco de la inducción automática de taxonomías y puede contribuir a solucionar el problema de la polisemia en hiperónimos de segundo nivel.

A continuación, se presenta un breve estado de la cuestión (apartado 2), la metodología (apartado 3), los resultados y evaluación (apartado 4) y las conclusiones y trabajo futuro (apartado 5). El código fuente del proyecto, implementado en el lenguaje Perl, se encuentra disponible en la página web que acompaña el artículo: http://www.tecling.com/hat

2 La asignación automática de hiperónimos

La hiperonimia es una de las relaciones semánticas de inclusión que acontecen en la estructura léxica de una lengua (García y Pascual, 2009). Leech (1985) la describió como el fenómeno por el cual una palabra incluye semánticamente a otra. Así, un hiperónimo se define como una unidad léxica cuyo significado está en un nivel de abstracción más alto que el de su hipónimo.

Una taxonomía léxica debe presentar las siguientes tres características fundamentales:

1. **Herencia**: un nodo inferior (hipónimo) hereda las propiedades de su nodo superior (hiperónimo).
2. **Asimetría**: una unidad léxica no puede ser superior (hiperónimo) e inferior (hipónimo) de otra unidad léxica al mismo tiempo.
3. **Transitividad**: si un hipónimo a tiene un hiperónimo directo ($a \rightarrow b$) y este, a su vez, tiene otro ($b \rightarrow c$), entonces el primero es hipónimo del último ($a \rightarrow c$).

En lingüística computacional se utilizaron términos como *red semántica* u *ontología* para referirse a estructuras de datos relacionados formalmente aplicadas al procesamiento automático de grandes cantidades de datos (Sowa, 2000). Cabe aclarar, sin embargo, que una taxonomía léxica es algo distinto a las anteriores estructuras, ya que solo establece relaciones de hiperonimia y además lo hace entre unidades léxicas, no entre conceptos. Una ontología no debería presentar problemas derivados de la polisemia porque puede identificar sus nodos conceptuales con un código arbitrario, tal como un identificador numérico. Esto le permite, además, asociar términos distintos para un mismo concepto, con lo cual se evita también el problema de la sinonimia, otra de las complicaciones de las taxonomías semasiológicas.

Los primeros intentos para construir taxonomías y ontologías se desarrollaron de forma manual, en casos como los de CyC (Lenat, 1995), WordNet (Fellbaum, 1998), EuroWordNet (Vossen, 2004), Snomed (Stearns et al., 2001), entre otros. Por supuesto, el desarrollo en forma manual de estas estructuras de datos presenta limitaciones. Por un lado,

son propensas a inconsistencias, incluso con protocolos rigurosos. Por otro lado, se vuelven obsoletas con relativa rapidez debido al dinamismo de la lengua, problema que se agudiza en el caso de las taxonomías especializadas, que tienen una acelerada evolución terminológica.

Estas limitaciones han sido motivo suficiente para emprender la tarea de la generación automática de taxonomías. Una primera línea de investigación consistió en la extracción de relaciones de hiperonimia a través del procesamiento automático de diccionarios (Calzolari, 1984; Chodorow, Byrd, y Heidorn, 1985; Guthrie et al., 1990; Agirre et al., 1994). Estos trabajos se basan fundamentalmente en la elaboración de sistemas de reglas que puedan analizar las definiciones y extraer pares hipónimo-hiperónimo. Una regla de este tipo puede ser que el primer sustantivo en la definición de un sustantivo será su hiperónimo.

Más tarde se aplicó una idea similar, es decir, la utilización de listas de patrones preestablecidos, no ya sobre diccionarios sino sobre texto libre (Hearst, 1992). Se realizaron múltiples variantes de este enfoque, tales como el intento de extraer estos patrones directamente del mismo corpus de manera inductiva (Snow, Jurafsky, y Ng, 2006).

En la actualidad, la inducción de taxonomías sigue enfrentando, entre otros, problemas de estructura y polisemia. En este marco, el presente estudio contribuye a mejorar los resultados de la inducción de taxonomías mediante una propuesta metodológica que se fundamenta en dos ámbitos: por una parte, en los principios de la semántica léxica, utilizando unidades léxicas de contextos sintagmáticos de los sustantivos en estudio; por otra parte, en la estadística de corpus fundada, en nuestro caso, en la aplicación de una medida de similitud distribucional (Grefenstette, 1994; Lin, 1998) que permitirá operacionalizar la similitud semántica entre palabras con el fin de obtener mediciones cuantificables y comprobables empíricamente.

3 Metodología

En esta sección detallamos la propuesta metodológica para asignar relaciones de hiperonimia de segundo nivel en los casos de polisemia, utilizando para ello una medida de similitud distribucional entre sustantivos. Como variable para la comparación, utilizamos los verbos con los que están asociados sintagmáticamente los sustantivos. Presentamos primero la versión más básica del método y luego una serie de variantes que van añadiendo complejidad.

La sección abre con la descripción del proceso de selección de la muestra para experimentación (3.1). Luego se presenta el método más básico, llamado *binario*, que utiliza vectores binarios y solo mide la frecuencia de coocurrencia entre verbos y sustantivos (3.2). A continuación, se describe el resto de las variantes del método: *ponderado*, que también utiliza vectores binarios pero con verbos seleccionados mediante una medida de asociación (3.3); *euclidiano*, que en lugar de vectores binarios utiliza números reales obtenidos a partir de la medida de asociación (3.4) y, finalmente, *dependencias*, que utiliza vectores binarios pero con verbos que se obtienen por relaciones de dependencia sintáctica (3.5).

3.1 Selección de la muestra para experimentación

Con el objeto de obtener tríadas como las mostradas en la Figura 1 (es decir, hipónimo + hiperónimo de nivel 1 + dos posibles hiperónimos de nivel 2), implementamos un script Perl que interroga la base de datos WordNet en castellano (Vossen, 2004). El script detecta la presencia de sustantivos en más de un synset, lo que puede ser interpretado como indicador de polisemia, y que muestren a la vez por lo menos un hipónimo. Esto permitió encontrar 26 casos que satisficieran el requerimiento de una frecuencia mínima de 100 ocurrencias en el corpus de trabajo (v. apartado 3.2.1). Como se puede ver en la Tabla 1, en cada tríada tenemos el hipónimo, que es el sustantivo objetivo (como *laucha*), el sustantivo polisémico es el hiperónimo de primer nivel (*ratón*) y el resto son los candidatos a hiperónimo de nivel 2 (*animal* o *artefacto*), entre los cuales el sistema debe elegir uno.

3.2 El método base: *binario*

3.2.1 Extracción de contextos de aparición de los sustantivos

Por cada sustantivo analizado tomamos una muestra de contextos de aparición. Para ello utilizamos el corpus esTenTen (Kilgarriff y Renau, 2013), versión 2011 (9.500 millones de palabras). Hicimos muestreos aleatorios de un máximo de hasta 5.000 concordancias por

Sust. objetivo	Hiper. nivel 1	Hiper. nivel 2
asaltante	ladrón	**humano** — artefacto
caniche	perro	**animal** — artefacto
chimpancé	mono	**animal** — prenda
laucha	ratón	**animal** — artefacto
tarántula	araña	**animal** — artefacto
...	...	...

Tabla 1: Ejemplos del tipo de tríada en estudio, con la opción correcta en negrita.

cada sustantivo, aunque en muchos casos la muestra fue menor debido a que no todos tienen tanta frecuencia de aparición en el corpus. Utilizamos una ventana de contexto de 10 palabras a derecha e izquierda, teniendo en cuenta la distancia variable en que se puede presentar la coocurrencia verbo-argumento y, en esta variante del método, nos limitamos a medir la frecuencia de coocurrencia. Para ello basta el etiquetado morfosintáctico del corpus EsTenTen. La Tabla 2 muestra un fragmento de contexto del sustantivo *tarántula* con el etiquetado del corpus.

Forma	Categoría gramatical	Lema
Si	CSUBX	si
una	ART	un
tarántula	NC	tarántula
pica	VLfin	picar
a	PREP	a
una	ART	un
persona	NC	persona
...	...	...

Tabla 2: Ejemplo de contexto de aparición del sustantivo *tarántula*.

3.2.2 Extracción de los verbos

Por cada uno de los sustantivos analizados, se recorrieron sus contextos de aparición registrando la frecuencia de los verbos con los que coocurren. De este modo, conservamos los verbos en los que se observa una frecuencia de coocurrencia de mínimo 5 casos, umbral arbitrario sobre el que es más improbable que la observación sea fruto de accidente o error.

3.2.3 Conformación de una matriz de coocurrencia con verbos

Una vez obtenidos los listados de coocurrencia sustantivo-verbo, se conformó una matriz $M_{i,j}$ en la que los sustantivos son dispuestos en las filas y los verbos con los que coocurren en las columnas. La Tabla 3 muestra la estructura de esta matriz. Uno de los sustantivos analizados, como *absolutismo*, coocurre con frecuencia con el verbo *abandonar*, lo que también sucede con el sustantivo *caniche* pero no así con *tarántula*.

	abandonar	*abarcar*	*abogar*	...
absolutismo	1	0	0	...
caniche	1	0	0	...
tarántula	0	0	0	...
...	...	...	...	...

Tabla 3: Ejemplificación de la matriz de coocurrencia sustantivo-verbo.

En esta variante del método optamos por valores binarios, como se muestra en la Tabla 3. El valor de la celda se define en (1), donde la frecuencia de coocurrencia $fr(i,j)$ debe superar un umbral u ($u = 5$).

$$M_{i,j} = \begin{cases} 1 & fr(i,j) > u) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

3.2.4 Formación de vectores clase para hiperónimos de segundo nivel

Los hiperónimos de segundo nivel utilizados en la muestra (por ejemplo *evento, animal, máquina*, etc.) resultan demasiado abstractos para crear un vector de coocurrencia directamente como se explica en el apartado 3.2.3. Esto motivó que la construcción de vectores se llevara a cabo de manera indirecta, a través de la suma de vectores de varios sustantivos pertenecientes a esas categorías. La Tabla 4 muestra algunos ejemplos de dicha selección, donde se ve la categoría y 10 sustantivos pertenecientes a ella. La selección de sustantivos es arbitraria, pero se trata en todos los casos de miembros prototípicos de cada categoría, y que tendrán frecuencia alta en el corpus.

Hiper. nivel 2	Hipónimos
animal	caballo, canario, canguro, delfín, elefante, gorrión, jirafa, león, lobo, ornitorrinco
máquina	aspiradora, automóvil, cocina, cortadora, estufa, horno, juguera, motocicleta, refrigerador, soldadora
prenda	blusa, calcetín, calzón, camisa, chaleco, cinturón, corbata, pantalón, polera, sudadera
...	...

Tabla 4: Ejemplo de construcción de vectores-clase.

Por cada sustantivo de estas categorías se extrajeron sus contextos de aparición tal co-

mo se describe en (3.2.1) y se construyó una matriz de coccurrencia sustantivo-verbo como en (3.2.3). La Tabla 5 muestra la forma en que se suman los vectores-miembro para obtener un vector-clase. Al igual que en la Tabla 4, los sustantivos se disponen en las filas y los verbos en las columnas. La diferencia aquí está en que estos sustantivos (H_i) son los diez miembros elegidos de cada categoría. La última fila, señalada con el símbolo VC, representa el vector-clase, y consiste en la suma de los vectores de cada uno de sus hipónimos. Esto significa que cada componente de VC tendrá valor 1 si existe al menos una celda con valor 1 en la columna correspondiente. De esta manera, por cada uno de esos tipos semánticos más abstractos, obtenemos un vector clase, representado por los verbos con los que coocurren sustantivos hipónimos de estos hiperónimos más abstractos.

	V_1	V_2	V_3	V_4	V_5	...	V_n
H_1	0	1	1	0	0	...	0
H_2	0	0	1	0	1	...	0
H_3	0	1	1	0	0	...	1
H_4	0	0	1	0	1	...	0
...	...	...	...	...	...	...	...
H_n	0	1	1	0	0	...	0
VC	0	1	1	0	1	...	1

Tabla 5: Esquematización de la suma de vectores para la conformación del vector-clase.

3.2.5 Cálculo de similitud entre vectores

Una vez poblada la matriz de los sustantivos objetivo (Tabla 3) y la de los vectores-clase (Tabla 5), el siguiente paso consiste en aplicar una medida de similitud entre el vector que corresponde a este sustantivo objetivo ($\vec{o}$) y cada uno de los vectores-clase ($\vec{VC}$) que representan a los hiperónimos de segundo nivel. Como medida de similitud aplicamos el índice de Jaccard (2), que es apropiado para la comparación de vectores binarios. Dados dos vectores A y B, la similitud se obtiene oponiendo la intersección a la unión.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

Así, en el caso de *tarántula*, la selección entre *animal* y *artefacto* (sus dos hiperónimos de segundo nivel) se realiza por medio de una función h según el valor de similitud recién explicado, tal como se muestra en la Ecuación 3, donde $\vec{o}$ puede ser *tarántula*, $\vec{VC}_k$ puede ser *animal* y $\vec{VC}_i$ *artefecto*. Siempre se elige una de las opciones.

$$h(\vec{o}) = \begin{cases} \vec{VC}_k & J(\vec{o},\vec{VC}_k) > J(\vec{o},\vec{VC}_i) \\ \vec{VC}_i & \text{otherwise} \end{cases} \tag{3}$$

3.3 Aplicación de una medida de asociación: la variante *ponderada*

Tal como anticipamos al comienzo de la sección, experimentamos con distintas variantes del método principal con el fin de contrastar resultados.

La variante *ponderada* del método es muy similar a la anterior, y sigue utilizando vectores binarios. La única diferencia es que ahora poblamos esos vectores mediante una mejor selección de los verbos, utilizando para ello una medida de asociación sintagmática (Ecuación 4). De forma similar a la Ecuación 1, solo tendrán valor 1 los pares sustantivo-verbo que tengan una ponderación mayor a un umbral mínimo que, a diferencia del caso anterior, ahora tiene otro valor ($u = 0{,}01$).

$$cooc(s,v) = \frac{f(s,v)}{\sqrt{f(s)}.\sqrt{f(v)}} \tag{4}$$

El resto del procedimiento es idéntico al método básico.

3.4 Uso de vectores con números reales en lugar de binarios: la variante *euclidiana*

Esta tercera variante del método está basada en la anterior (*ponderada*), pero en lugar de utilizar valores binarios ahora son vectores de números reales, cuyos valores se obtienen de la ponderación definida en la Ecuación 4. El uso de valores reales en lugar de binarios obliga a hacer ajustes en el método básico, como la conformación de los vectores-clase (apartado 3.2.4). En este caso, los valores de los vectores-clase se obtienen sumando las ponderaciones de los verbos asociados a cada sustantivo, como se indica en la Ecuación 5.

$$VC_j = \sum_{i=1}^{|VC|} H_{i,j} \tag{5}$$

Otra de las diferencias en esta variante del método es que al utilizar vectores con números reales podemos optar por otras medidas

de similitud. En este caso optamos por la utilización de la distancia euclidiana, definida en la Ecuación 6.

$$d(A,B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} \qquad (6)$$

Esta variante permite captar la idea según la cual un sustantivo objetivo *o* y un hiperónimo de segundo nivel VC_k correcto deberían tener un similar perfil de coocurrencia con verbos, siendo algunos verbos más significativos que otros. Con esta medida se selecciona un hipónimo de la misma forma que en la Ecuación 3, solo que en este caso el signo $<$ se invierte a $>$, ya que se trata de una medida de distancia en lugar de similitud.

3.5 Uso de un *parser* sintáctico para extraer relaciones verbo-sustantivo: la variante *dependencias*

La última variante, y la más compleja, involucra la utilización de un analizador de dependencias sintácticas para determinar la función gramatical que se produce entre sustantivos y verbos. El *parser* permite limitar la selección a las parejas sustantivo-verbo que efectivamente contraen una relación sintáctica, como puede ser el caso de la relación sujeto-verbo, verbo-objeto directo, etc.

Utilizamos para ello UDPipe, uno de los mejores y más recientes analizadores de dependencias (Straka y Straková, 2017). La Tabla 6 muestra el resultado del análisis sintáctico del mismo fragmento de contexto del sustantivo *tarántula* que se mostró en la Tabla 2. En este caso se produce un error, ya que el verbo *picar* no es reconocido como tal (línea 96) y se etiqueta como adjetivo, perdiéndose así la información relativa a que *tarántula* es el sujeto del verbo *picar*. Dada la mayor complejidad de este tipo de análisis, cabe esperar que se produzca una alta tasa de error. Sin embargo, al mismo tiempo resulta razonable suponer también que la gran cantidad de contextos de aparición de los sustantivos compense esta tasa de error.

En este caso optamos nuevamente por la utilización de vectores binarios, pero no aplicamos un límite de frecuencia por ser ya esta variante muy selectiva. De este modo, si se observa al menos una vez que existe una relación sintáctica entre un sustantivo y un verbo, el valor correspondiente a esa celda será 1.

Línea	Forma	Lema	POS	Dep.	Func.
93	Si	si	SCONJ	102	mark
94	una	uno	DET	95	det
95	tarántula	tarántula	NOUN	102	nsubj
96	pica	pico	ADJ	95	amod
97	a	a	ADP	99	case
98	una	uno	DET	99	det
99	persona	persona	NOUN	95	nmod
...	...	...	...	...	...

Tabla 6: Ejemplo de análisis de dependencias con *UDPipe* en que se produce un error en la detección del verbo *picar*.

4 Resultados

La Tabla 7 muestra los resultados del método en sus distintas variantes, sobre la muestra de sustantivos objetivo. La cobertura es igual a la precisión debido a que forzamos al sistema a elegir siempre una de las opciones.

Variante	Precisión
binaria	69 %
ponderada	**84 %**
euclidiana	73 %
dependencias	57 %
aleatoria	50 %

Tabla 7: Resultados del método en sus distintas variantes

Los mejores resultados se obtuvieron con la variante *ponderada*. Atribuimos este resultado a una mejor selección de los verbos coocurrentes, a través de una medida de asociación sintagmática. Esto suprimió el ruido que introducían en la variante *binaria* los verbos que tienen alta frecuencia de coocurrencia con muchos sustantivos diferentes. La variante *euclidiana*, con la incorporación de la medida de distancia euclidiana, también mejora la variante *binaria*, pero con resultados más modestos. Finalmente, la variante *dependencias*, que extrae los verbos mediante análisis sintáctico y es la de mayor complejidad, tuvo el peor desempeño, con solo 7 puntos por encima de una clasificación aleatoria. Esto puede ser atribuible al hecho de que los textos del corpus, tomados de páginas web, contienen una sintaxis y ortografía relajadas propias de los textos de Internet. Solucionar este problema queda fuera del alcance de la presente investigación.

La Tabla 8 muestra los resultados con las primeras 9 unidades analizadas por orden alfabético en el caso de la variante *ponderada*, que fue la que produjo mejores resultados. La primera columna indica la evaluación: 1 si el ensayo es exitoso y 0 si no lo es. La si-

E	d	h_1	h_2	S
1	absolutismo	sistema	concepto	16.46
			máquina	8.47
0	acueducto	canal	institución	11.55
			lugar	8.12
0	albahaca	planta	lugar	14.15
			servivo	11.09
1	ametrallador	cañón	arma	5.56
			lugar	2.58
1	asaltante	ladrón	humano	11.46
			artefacto	8.07
1	caniche	perro	animal	6.26
			artefacto	3.84
1	cencerro	campana	instr. musical	3.07
			máquina	2.97
1	chimpancé	mono	animal	10.47
			prenda	5.75
1	dedo	miembro	parte cuerpo	11.39
			humano	10.92

Tabla 8: Ejemplos de resultados con la variante *ponderada*

guiente columna presenta el sustantivo objetivo (o), la siguiente el hiperónimo de primer nivel ($h1$) y la siguiente los distintos hiperónimos de segundo nivel ($h2$). La última columna (S) indica el valor obtenido con el índice de Jaccard entre el vector de coocurrencia del sustantivo objetivo y, en cada caso, el hiperónimo de segundo nivel (el valor más alto se presenta primero). Entre los 26 casos estudiados hay 22 exitosos, lo que representa un resultado estadísticamente significativo ($p = 0{,}0005$).

5 Conclusiones y trabajo futuro

En este trabajo hemos presentado una propuesta metodológica para desambiguar la relación entre un sustantivo y un hiperónimo polisémico en el contexto de la inducción automática de taxonomías. El método, fundado en una medida de similitud distribucional, se basa en la idea según la cual las palabras que aparecen en cotextos similares tienden a tener significados similares.

En la propuesta, se han restringido los cotextos en función de los verbos con los que coocurren los sustantivos en estudio, ya que se trata de una clase abierta de palabras, pero al mismo tiempo limitada (cerca de 6.000 verbos aparecen con cierta frecuencia en el EsTenTen). Esta característica convierte a los verbos en predictores útiles para obtener información semántica acerca de los sustantivos con los que coocurren, ya que representan una matriz más manejable que una de adjetivos o sustantivos.

Además del método básico, hemos explorado distintas variantes. Utilizamos vectores binarios que indican la coocurrencia sustantivo-verbo, y también vectores con valores reales; probamos el uso de simple frecuencia de coocurrencia y luego una medida de asociación estadística y, finalmente, hemos explorado también la posibilidad de extraer los verbos por medio de un analizador de dependencias sintácticas.

Nuestros resultados permiten concluir que la mejor variante es la que utiliza vectores binarios que miden frecuencia de coocurrencia sustantivo-verbo, seleccionando los verbos con una medida de asociación (la variante *ponderado*). Creemos que la tasa de éxito es remarcable, teniendo en cuenta que se trata de un método relativamente simple. No existen, que sepamos, propuestas similares para la adjudicación de hiperónimos de segundo nivel en casos de polisemia.

En cuanto a trabajo futuro, es necesario continuar introduciendo nuevas variantes metodológicas y reproducir los experimentos con muestras más grandes de datos, ya que esto permitiría estudiar mejor cómo afectan a los resultados las diferencias de frecuencia y de prototipicidad de cada significado, una variable que conviene controlar en un diseño de investigación de este tipo (por ejemplo, en el caso del sustantivo *perro*, el sentido de *animal* tendrá más peso que el de *artefacto*). También sería necesario probar la utilización de ventanas oracionales en lugar de ventanas de contexto simétricas. Otra posibilidad sería reproducir el mismo método pero utilizando adjetivos o incluso sustantivos en lugar de verbos. Finalmente, proyectamos reproducir los experimentos en otras lenguas (francés, inglés, etc.) en el contexto de nuestro proyecto KIND[1] de taxonomías automatizadas en varias lenguas.

Agradecimientos

Esta investigación ha sido posible gracias al financiamiento del Proyecto Fondecyt Regular 1191204 “Polisemia regular de los sustantivos del español: análisis semiautomático de corpus, caracterización y tipología”, dirigido por Irene Renau. Agradecemos también a los revisores por sus útiles comentarios.

[1] http://www.tecling.com/kind

Bibliografía

Agirre, E., X. Arregi, X. Artola, A. D. de Ilarraza, y K. Sarasola. 1994. A methodology for the extraction of semantic knowledge from dictionaries using phrasal patterns. En *Proceedings of IBERAMIA'94. IV Congreso Iberoamericano de Inteligencia Artificial*, páginas 263–270, Caracas (Venezuela).

Baldinger, K. 1977. *Teoría semántica: hacia una semántica moderna*. Coleccion Romania. Alcala.

Bordea, G., P. Buitelaar, S. Faralli, y R. Navigli. 2015. SemEval-2015 Task 17: Taxonomy extraction evaluation (texeval). En *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, páginas 902–910. ACL.

Bordea, G., E. Lefever, y P. Buitelaar. 2016. SemEval-2016 Task 13: Taxonomy extraction evaluation (texeval-2). En *SemEval-2016*, páginas 1081–1091. ACL.

Calzolari, N. 1984. Detecting patterns in a lexical data base. En *Proceedings of the 10th International Conference on Computational Linguistics and 22nd annual meeting on ACL*, páginas 170–3. ACL.

Chodorow, M. S., R. J. Byrd, y G. E. Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. En *Proceedings of the 23rd annual meeting on ACL*, páginas 299–304. ACL.

De Miguel, E. 2016. Lexicología. En J. Gutiérrez, editor, *Enciclopedia de Lingüística Hispánica*. Ariel, Barcelona, páginas 153–185.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

García, R. y J. Pascual. 2009. Relaciones de significado entre las palabras. En E. D. Miguel, editor, *Panorama de lexicología*. Ariel, Barcelona, páginas 117–131.

Grefenstette, G. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA, USA.

Guthrie, L., B. Slator, Y. Wilks, y R. Bruce. 1990. Is there content in empty heads? En *Proc. of the 13th International Conference on Computational Linguistics, COLING'90 (Helsinki, Finland)*, páginas 138–143.

Hearst, M. A. 1992. Automatic acquisition of hyponyms from large text corpora. En *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, páginas 539–545, Stroudsburg, PA, USA. ACL.

Kilgarriff, A. 1992. *Polisemy*. Ph.D. tesis. University of Sussex.

Kilgarriff, A. y I. Renau. 2013. estenten, a vast web corpus of peninsular and american spanish. *Procedia - Social and Behavioral Sciences*, 95:12 – 19.

Klapaftis, I. P. y S. Manandhar. 2010. Taxonomy learning using word sense induction. En *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, páginas 82–90, Los Angeles, California, Junio. ACL.

Leech, G. 1985. *Semántica*. Alianza Universal, No. 197. Alianza.

Lenat, D. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11):33–38, Noviembre.

Lin, D. 1998. Automatic retrieval and clustering of similar words. En *Proceedings of the 17th International Conference on Computational Linguistics - Volume 2*, COLING '98, páginas 768–774, Stroudsburg, PA, USA. ACL.

Lyons, J. 1977. *Semantics*, volumen 2. Cambridge University Press.

Sager, J. C. 1990. *A Practical Course in Terminology Processing*. John Benjamins, Amsterdam/Philadelphia.

Snow, R., D. Jurafsky, y A. Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. En *Proceedings of the 21st International Conference on Computational Linguistics, Sydney, Australia, 17-21 July 2006*.

Stearns, M. Q., C. Price, K. A. Spackman, y A. Y. Wang. 2001. Snomed clinical terms: overview of the development process and project status. En *Proceedings of the AMIA Symposium*, páginas 662–666. American Medical Informatics Association.

Straka, M. y J. Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. En *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, páginas 88–99, Vancouver, Canada, Agosto. ACL.

Ullmann, S. 1972. *Semántica*. Aguilar.

Velardi, P., S. Faralli, y R. Navigli. 2013. Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics*, 39(3):665–707.

Vossen, P. 2004. Eurowordnet: A multilingual database of autonomous and language-specific wordnets connected via an inter-lingual-index. *Special Issue on Multilingual Databases, International Journal of Linguistics*, 17(2):161–173, 06.