Revista Signos ISSN 0718-0934
© 2011 PUCV, Chile
DOI: 10.4067/S0718-09342011000100005

44(75) 68-84

Hacia la identificación de relaciones de hiponimia/hiperonimia en Internet*

Towards the identification of hyponym/hypernym relations in the Internet

Rosa María Ortega
mortega@itesa.edu.mx
Instituto Superior del Oriente del Estado de Hidalgo
México

César Aguilar
caguilar@iingen.unam.mx
Universidad Autónoma de Querétaro
México

Luis Villaseñor
villasen@inaoep.mx
Instituto Nacional de Astrofísica, Óptica y Electrónica
México

Manuel Montes
mmontesg@inaoep.mx
Instituto Nacional de Astrofísica, Óptica y Electrónica
México

Gerardo Sierra
gsierram@iingen.unam.mx
Universidad Nacional Autónoma de México
México

Recibido: 2-VII-2009 / Aceptado: 30-XI-2010

Resumen: En este trabajo se presenta un enfoque para la extracción automática de pares hipónimo-hiperónimo. En particular se propone un método de extracción de información léxica, orientado a la relación de hiponimia, que utiliza un conjunto de patrones léxicos propios del español, así como un esquema simétrico de calificación de pares/patrones cuyo objetivo es enriquecer la confiabilidad del método de extracción. La eficacia del método propuesto se evaluó obteniendo hipónimos correspondientes a un vocabulario de hiperónimos dado. Los resultados logrados confirman la utilidad del método propuesto para extraer hipónimos, así como la relevancia del esquema de calificación de pares/patrones.

Palabras Clave: Hipónimo, hiperónimo, patrones léxico-sintácticos, extracción de información.

Abstract: This paper presents an approach to the automatic extraction of hyponyms and hyperonyms. In particular, it proposes an information extraction method that is specially suited for identifying pairs of hyponym-hyperonym by using a set of Spanish lexical patterns. It also proposes a symmetric weighting scheme of pairs/patterns whose goal is to enhance the confidence of the extraction method. The effectiveness of the proposed approach was evaluated by extracting hyponyms from a given vocabulary of hyperonyms. Results show the usefulness of the proposed extraction method as well as the relevance of the pairs/patterns weighting scheme.

Key Words: Hyponym, hypernym, lexical-syntactic pattern, information extraction.

INTRODUCCIÓN

La búsqueda y extracción de información en Internet juega un papel relevante para la lexicografía y la terminología actuales, lo que ha llevado a implementar nuevos métodos y técnicas para acceder a esta información (Llisterri, 2003; Águila, 2006; Rojo, 2008). Muchos de estos métodos y técnicas son híbridos, pues emplean el conocimiento aportado por la lingüística, la estadística y las ciencias computacionales para resolver tareas como la construcción de diccionarios electrónicos (Wilks, Slator & Guthrie, 1996), terminologías (Cabré, Estopà & Vivaldi, 2001) o redes léxicas (Fellbaum, 1998), por mencionar algunos recursos relevantes.

El presente trabajo se sitúa en el terreno de la extracción de información y su objetivo principal consiste en delimitar un método de extracción de pares hipónimo/hiperónimo usando un conjunto de patrones léxicos propios del español. Básicamente, el método propuesto en esta investigación aplica dichos patrones a documentos recopilados de Internet (textos escritos en prosa) y detecta automáticamente un conjunto de hipónimos relacionados a un vocabulario previamente proporcionado.

Siguiendo el enfoque planteado por Hearst (1992), consideramos el uso de patrones léxico-sintácticos para llevar acabo nuestro proceso de extracción. Este enfoque parte de la idea de que en una lengua existe esta clase de patrones, los cuales permiten expresar hipónimos e hiperónimos dentro de un texto. Por ejemplo, la frase 'es un' es comúnmente utilizada como un operador que relaciona un hipónimo con su respectivo hiperónimo (Wilks et al., 1996).

Tales patrones, como mostró Hearst (1992), pueden ser aplicados dentro un proceso de búsqueda automática para recuperar hipónimos de una colección de textos. A la fecha existen varias propuestas para descubrir patrones de hiponimia de manera automática (Pasca, 2004; Pantel & Pennacchiotti, 2006; Barbu, 2008). En particular, en esta investigación se utiliza el conjunto de patrones que han sido recuperados a través del método desarrollado por Ortega (2007).

Naturalmente, dada la riqueza y complejidad que conlleva toda lengua humana, los enfoques automáticos basados en patrones para descubrir relaciones léxicas no son completamente confiables. Es por ello, que para mejorar su precisión se evalúa la confianza de los patrones encontrados y/o la confianza de los pares hipónimo-hiperónimo detectados (Pantel & Pennacchiotti, 2006; Ortega, Villaseñor & Montes, 2007; Blohm, Cimiano & Stemle, 2007; Barbu, 2008).

En esta investigación, evaluamos la calidad de los pares extraídos al estimar un valor de confianza de los patrones aplicados para extraerlos. De manera simétrica, también la calidad de los patrones es estimada mediante un valor de confianza de los pares extraídos. Un esquema similar lo presentan Pantel y Pennacchiotti (2006), donde se utiliza una medida probabilística, 'la información mutua', para medir el grado de asociación entre pares y patrones. En contraste, en el presente estudio, 'la medida F', utilizada tradicionalmente en el área de recuperación de información y propuesta por Van Rijsbergen (1979) es adaptada para calcular un valor de confianza de los patrones, así como de los

pares descubiertos. Precisamente, este esquema simétrico de calificación pares/patrones es una de las contribuciones principales de esta investigación. Gracias a la integración de este esquema es posible determinar con mayor precisión los pares de hipónimos/hiperónimos.

Antes de explicar en detalle el método propuesto en esta investigación, las siguientes secciones aportan algunos antecedentes sobre las relaciones de hiponimia e hiperonimia, incluyendo el uso de patrones para su extracción. Posteriormente, en las secciones finales, se discuten los resultados alcanzados y se presentan las conclusiones de este trabajo.

1.Relaciones de hiponimia e hiperonimia

Se denomina hiperónimo a aquel término general que puede ser utilizado para referirse a la realidad nombrada por un término más particular o hipónimo. Así, un hiperónimo no posee ningún rasgo semántico, que no comparta su hipónimo, mientras que éste sí posee rasgos semánticos que lo diferencian de aquél. En otras palabras, el significado del concepto más específico (hipónimo) está incluido en el significado del concepto más general (hiperónimo) (Cruse, 1986). Ejemplos de pares hipónimo/hiperónimo son los siguientes:

i. Gorrión [hipónimo]pájaro [hiperónimo]
ii. Pájaro [hipónimo] animal [hiperónimo]
iii. Animal [hipónimo] entidad [hiperónimo]

Donde los rasgos semánticos de 'animal' son compartidos por los de 'pájaro', pero este posee otros rasgos que lo diferencian del primero. Dicha relación de inclusión, en un plano léxico, permite establecer clasificaciones y jerarquías, de modo que puede hacerse patente cómo se relaciona conceptualmente una palabra con otras.

De acuerdo con Cruse (1986), Wilks et al. (1996), así como Miller (1998), las relaciones de hiponimia e hiperonimia son aquellas que se dan, dentro de un plano léxico-semántico, entre dos o más palabras, de tal suerte que una de ellas se subordina conceptualmente a otra. Esto equivale a decir que el concepto referido por una palabra (hipónimo) es una instancia concreta de un objeto prototípico (hiperónimo), situado jerárquicamente en un nivel superior.

Siguiendo con la explicación hecha por Miller (1998), las relaciones de hiponimia e hiperonimia son básicas dentro de toda interfaz léxico-semántica de una lengua natural, debido a que una de sus funciones más importantes es estructurar sistemas de conceptos dentro de la mente de un humano, organizados conforme a las propiedades o atributos que tales conceptos prediquen de una entidad o un evento. Si bien esta clase de información es reconocible en cualquier palabra, áreas de estudio como la lexicografía, la lexicografía computacional o la extracción de información han caracterizado a los nombres como unidades léxicas prototípicas que proyectan vínculos de hiponimia e hiperonimia.

Dentro del procesamiento del lenguaje natural, el mejor ejemplo de cómo se han explotado estos vínculos entre nombres ha sido la creación de la red léxica conocida como *Wordnet* (Fellbaum, 1998), la cual es justo un sistema jerárquico de clasificación automático, el cual permite asociar nombres como hipónimos o hiperónimos entre sí, de suerte que puede visualizarse cuáles son los nexos conceptuales que mantienen tales nombres.

1.1. Aplicaciones de relaciones de hiponimia e hiperonimia

Dentro del campo de la ingeniería lingüística, las relaciones de hiponimia e hiperonimia se han usado en tres áreas específicas:

a) En la creación de diccionarios y otros recursos de consulta léxica, cuya información proviene de repositorios textuales. Un trabajo representativo en esta área es el de Wilks et al. (1996), orientado hacia la extracción de definiciones, el cual ha dado lugar a varias propuestas. En español, cabe mencionar la de Denicia, Montes, Villaseñor y García (2006), o la de Sierra, Alarcón, Aguilar y Bach (2008).

b)En el diseño de sistemas para la detección de unidades textuales cuya información léxica aporte un conocimiento determinado (nombres de personas, términos, eventos, etc.), e igualmente ayuden a

desambiguar el sentido de una palabra en un contexto dado. Trabajos como los de Hearts (1992), Girju, Badulescu y Moldovan (2006), así como Snow, Jurafsky y Ng (2006) se ubican en esta área.

c) Finalmente, en el desarrollo de taxonomías y ontologías, cuya estructuración se basa precisamente en tales relaciones. Como ejemplos de esta clase de desarrollos se encuentra el trabajo de Snow et al. (2006) o el manual editado por Buitelaar, Cimiano y Magnini (2007).

Es importante notar la relación existente entre las áreas mencionadas. Generalmente, la creación automática de recursos léxicos para un idioma o dominio específico corresponde al resultado de la integración de métodos automáticos que extraen relaciones léxicas entre dos entidades. Tomando en cuenta esta observación, Buitelaar et al. (2007) han elaborado un manual sumamente completo en donde exponen, discuten y evalúan varios algoritmos para extraer estas relaciones, junto con métodos que estructuran la información encontrada para crear repositorios semánticos.

Por otro lado, los mismos autores mencionados consideran la relación de hiponimia como la columna vertebral de las ontologías, ya que permite estructurar conceptos en categorías semánticas. Así, el texto editado por Buitelaar et al. (2007) ofrece una interesante perspectiva sobre cómo aprovechar métodos automáticos que extraigan instancias de la relación de hiponimia, con miras a dar un primer paso sólido hacia la construcción automática de ontologías.

Las siguientes secciones describen el método propuesto para extraer pares hipónimo/hipéronimo, detallando el proceso de evaluación de pares hipónimo-hiperónimo; así como también, los resultados experimentales.

2. Extracción de pares hipónimo/ hiperónimo

Dentro del campo de la extracción de información (EI) -vista como un área interdisciplinaria enfocada en la identificación automática de unidades textuales con información-, se han desarrollado sistemas automáticos capaces de identificar y extraer relaciones léxicas en grandes repositorios de documentos e Internet. Por mencionar algunos, Baroni y Bisi (2004) reportan un sistema para extraer relaciones de sinonimia, mientras que Lucero, Pinto y Jiménez (2004) estudian la extracción de antónimos. Finalmente, Pennacchiotti y Pantel (2009) extraen distinta información léxica como nombres de actores, atletas y músicos.

Los sistemas de extracción tratan de hacer inferencias para detectar relaciones léxicas, basándose en el uso de patrones léxicos y sintácticos (Gelbukh & Sidorov, 2006). Respecto al caso de la extracción de hipónimos e hiperónimos, se ha observado que existen patrones léxico-sintácticos que codifican esta relación. Hearts (1992), pionera en el uso de patrones para extraer relaciones semánticas, reporta la serie de patrones mostrados en la Tabla 1. Para mayor claridad, en la Tabla 1 también se muestran los patrones traducidos al español. En la traducción se consideró la variedad en número (singular o plural) abstraída en la frase nominal (FN) de un enunciado en inglés.

Tabla 1. Lista de patrones propuestos por Hearts (1992: 541).

Patrones en inglés	Patrones traducidos al español
FN *such as* {FN, FN… *(and/or)*} FN	FN *tal(es) como* {FN, FN… *(y/o)*} FN
Such FN *as* {FN, }* {*(and/or)*} FN	*Tal(es)* FN *como* {FN, }* {*(y/o)*} FN
FN {, FN}* {,} *or other* FN	FN {, FN}* {,} *u otro(s)* FN
FN {, FN}* {,} *and other* FN	FN {, FN}* {,} *y otro(s)* FN
FN {,} *including* {FN ,} * {*and/or*} FN	FN {,} *incluyendo* {FN ,} * {*y/o*} FN
FN {,} *especially* {FN ,} * {*and/or*} FN	FN {,} *especialmente* {FN ,} * {*y/o*} FN

Este tipo de patrones permite reconocer pares de palabras situadas en una relación de hiponimia/ hiperonimia, por ejemplo:

Works by such ***authors*** *as* Herrick, Goldsmith *and* Shakespeare.

Aquí, los nombres Herrick, Goldsmith y Shakespeare son reconocidos como hipónimos de *authors,* gracias al patrón léxico-sintáctico FN *such as* {*FN, FN… (and/or)*} *FN.*

El presente trabajo, retoma la propuesta de Hearst (1992) y presenta los resultados obtenidos a partir de un experimento para el español, el cual consistió en la extracción de relaciones de hiponimia/ hiperonimia utilizando documentos recuperados de Internet. Cabe resaltar que a diferencia del trabajo de Hearst (1992), los patrones usados son más simples, limitándose a usar únicamente elementos léxicos.

3.La investigación

La secuencia que se seguirá para exponer este trabajo de investigación será la siguiente: en primer término, se establecen los patrones léxicos en español que codifican la relación de hiponimia. Posteriormente, se presenta la arquitectura diseñada para la extracción automática de hipónimos.

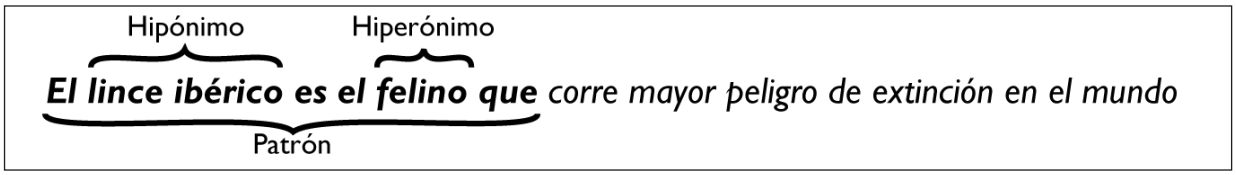
3.1. Patrones léxicos

El trabajo de Hearst (1992) ha sido la base de varias propuestas que se hacen uso de patrones para extraer hipónimos. Existen trabajos que buscan patrones automáticamente y otros que los desarrollan manualmente. En nuestro caso, nuestra propuesta utiliza la serie de patrones descubiertos automáticamente por Ortega (2007). Tales patrones se muestran en la Tabla 2.

Brevemente, estos patrones son obtenidos a partir de pares elegidos manualmente representando una relación de hiponimia como el par 'águila-ave'. Estos pares, los cuales reciben el nombre de 'semillas', son utilizados para recuperar ejemplos, es decir, fragmentos de texto, que revelan cómo las personas relacionan textualmente un hipónimo con su hiperónimo. Naturalmente, es necesario contar con ejemplos verídicos.

Tabla 2. Lista de patrones utilizados, tomado de Ortega (2007).

No.	Patrón
1	el <hipónimo> es el único <hiperónimo>
2	el uso de la <hipónimo> como <hiperónimo>
3	el <hipónimo> es uno de los <hiperónimo> más
4	de la <hipónimo> como <hiperónimo> de
5	de las <hipónimo> como <hiperónimo>
6	las <hipónimo> son una <hiperónimo>
7	el <hipónimo> es un <hiperónimo> que
8	el <hipónimo> es el <hiperónimo> que
9	de <hiperónimo> como <hipónimo> y
10	la <hipónimo> es un <hiperónimo>
11	la <hipónimo> una <hiperónimo>
12	las <hipónimo> son <hiperónimo> que
13	el <hipónimo> es un <hiperónimo> de
14	la <hipónimo> es la <hiperónimo>
15	la <hipónimo> es una <hiperónimo> que
16	la <hipónimo> como una <hiperónimo>
17	que la <hipónimo> es una <hiperónimo>
18	el <hipónimo> es una <hiperónimo>
19	la <hipónimo> es el <hiperónimo> de
20	de <hipónimo> y otras <hiperónimo>
21	del <hipónimo> como <hiperónimo>
22	el <hipónimo> es la <hiperónimo>
23	<hiperónimo> de <hipónimo> de
24	de <hipónimo> y <hiperónimo>
25	<hiperónimo> de <hipónimo> y
26	de <hipónimo> o <hiperónimo>
27	los <hipónimo> son <hiperónimo>
28	de <hipónimo> como <hiperónimo> de
29	el <hipónimo> y las <hiperónimo>
30	de los <hipónimo> y <hiperónimo>
31	de los <hipónimo> y los <hiperónimo>
32	la <hipónimo> es el único <hiperónimo> natural
33	<hiperónimo> de la actividad <hipónimo> y el deporte
34	la anorexia y la <hipónimo> son <hiperónimo>
35	de <hipónimo> y otros <hiperónimo>
36	el <hipónimo> es el <hiperónimo> de mayor longevidad
37	los <hipónimo> y otros <hiperónimo>
38	facultad de <hiperónimo> de la actividad <hipónimo> y
39	la <hipónimo> y otros <hiperónimo>
40	las <hipónimo> marinas son <hiperónimo>
41	el <hipónimo> es el <hiperónimo> interno más
42	licenciado en <hiperónimo> de la actividad <hipónimo> y del deporte
43	el <hipónimo> es el <hiperónimo> más grande del cuerpo



Cuadro 1. Fragmento textual con un par seleccionado (Ortega, 2007: 55).

Posteriormente, los ejemplos recopilados son procesados por una técnica de minería de texto, la cual permite extraer las secuencias frecuentes maximales (SFM's) (Ahonen-Myka, 2002). Específicamente, para obtener las SFM's se utilizó la implementación presentada en (García-Hernández et al., 2006). Lo anterior, con el fin de generalizar las convenciones o frases que las personas utilizan para introducir una relación de hiponimia. Las frases resultantes representan los patrones de extracción de pares hipónimo/hiperónimo. Gracias a estos patrones, es posible localizar, en fragmentos de documentos recuperados de Internet, pares hipónimo/hiperónimo, tal como se muestra en el Cuadro 1.

En este caso, el sistema detecta el par conformado por las palabras 'lince ibérico' y 'felino' a través del patrón número 8 de la Tabla 2. Como es posible imaginar, la aplicación de un patrón no siempre resulta en la detección de un par hipónimo/hiperónimo correcto. El Cuadro 2 muestra cómo el mismo patrón 8, utilizado en el Cuadro 1, no detecta un par correcto. De ahí que sea necesario aplicar otro proceso para identificar a los pares correctos. Con base en lo reportado en (Ortega et al., 2007), el método propuesto estima un valor de confianza para cada par detectado, y reporta como resultado final el conjunto de aquellos pares que superan cierto umbral de confianza preestablecido. A continuación se detalla la arquitectura del método propuesto.

3.2. Arquitectura del sistema de extracción

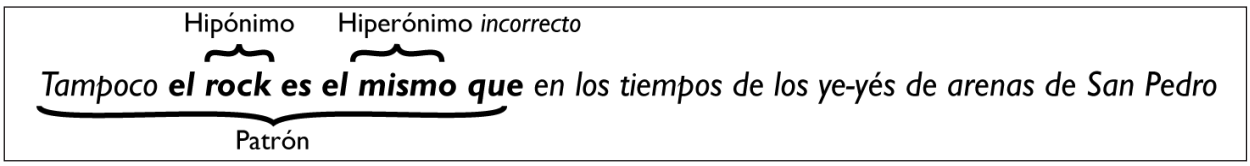
El sistema propuesto consiste de dos etapas para la detección de relaciones de hiponimia/hiperonimia en textos recuperados de Internet:

- En la primera etapa se construye un catálogo de posibles pares hipónimo/hiperónimo. Este se realiza formulando una serie de consultas en un buscador Web para ubicar fragmentos de texto donde puedan aparecer pares hipónimo/hiperónimo.
- En la segunda etapa, se estima un valor de confianza para cada par hipónimo/hiperónimo detectado. Como resultado de esta segunda etapa, es posible ordenar el catálogo gracias a los valores de confianzas estimados.

Las siguientes secciones describen estas dos etapas.

3.2.1. Construcción del catálogo de posibles pares hipónimo/hiperónimo

La finalidad de esta etapa es recopilar un catálogo de hipónimos para un vocabulario predefinido. Básicamente, el vocabulario es una lista de hiperónimos para los cuales se recuperarán hipónimos. La recopilación se lleva acabo usando un buscador *Web*. En específico, se utilizó el buscador *Google*. La Figura 1 muestra los pasos en la construcción de este catálogo.



Cuadro 2. Detección de un par hipónimo/hiperónimo incorrecto.

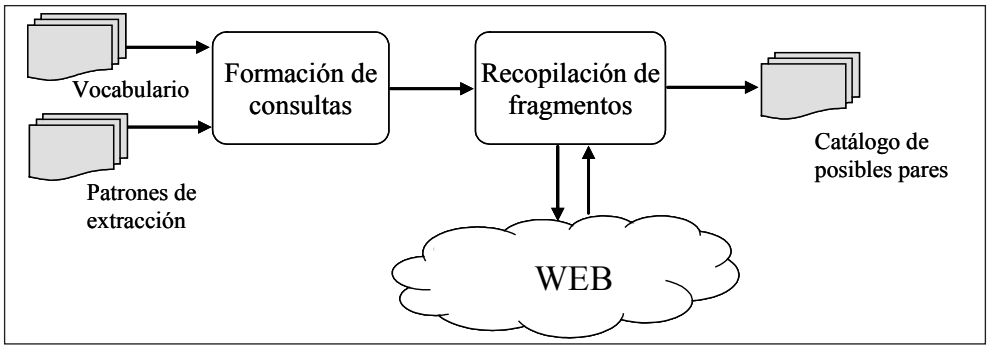


Figura 1. Construcción del catálogo de posibles pares hipónimo/hiperónimo.

El primer paso de esta etapa consiste en formar consultas que se entregarán a dicho buscador. Para formar las consultas es necesario 'aterrizar' cada uno de los patrones con los términos del vocabulario. Dicho de otra manera, cada uno de los términos del vocabulario substituirá la etiqueta <hiperónimo> de cada patrón. Entonces, la etiqueta <hipónimo> actuará como un comodín extrayendo posibles hipónimos para los términos del vocabulario. Para ilustrar este proceso, considere la palabra felino como un término del vocabulario y el patrón 1 de la Tabla 2: el <hipónimo> es el único <hiperónimo>. El patrón de consulta formado corresponde al siguiente: el <hipónimo> es el único felino, donde las palabras 'el' y 'es' delimitarán las unidades textuales que representen un posible hipónimo.

Enseguida, cada consulta es utilizada para recopilar fragmentos de texto asociados a dicha consulta. A los fragmentos recuperados se les aplica el patrón que les dio origen, con el fin de detectar un posible par hipónimo/hiperónimo con el cual se crea una entrada en el catálogo. Por ejemplo, retomando el extracto del Cuadro 1, la entrada en el catálogo corresponde a 'lince ibérico – felino'.

Naturalmente, dado que es un método automático, no todas las entradas en el catálogo serán correctas. De ahí, que lo llamemos catálogo de 'posibles' pares hipónimo/hiperónimo. Más específicamente, existen principalmente dos aspectos que propician esta situación. Primero, la variedad lingüística de un idioma necesita un conjunto amplio de patrones para abstraer una relación de hiponimia. Si bien, es una lista extensa, no se considera un conjunto exhaustivo que generalice todas las frases que pueden introducir una relación de hiponimia en los textos. Segundo, los patrones son descubiertos automáticamente y aun cuando tengamos una lista completa de patrones, habrá situaciones donde el patrón no extraiga un par correcto, tal es el caso de la aplicación del patrón 8 en el Cuadro 2.

Dadas estas situaciones, se propone una segunda etapa donde se estime el valor de confianza para cada par en el catálogo. De esta manera, se seleccionan aquellos pares con mayor probabilidad de ser correctos.

3.2.2. Estimación del valor de confianza de los pares hipónimo/hiperónimo

Para estimar el valor de confianza de un par hipónimo/hiperónimo, el método propuesto utiliza un proceso auto-sustentado basado en dos supuestos: (i) un patrón es más confiable mientras mayor sea la cantidad de pares confiables detectados por el mismo; y (ii) un par es más confiable mientras mayor sea la cantidad de patrones que lo detectan. Bajo estos supuestos, se sugiere un esquema donde el valor de confianza de los patrones ayude a estimar el valor de confianza de los pares, y asimismo, que el valor de confianza de los pares ayude a determinar el valor de confianza de los patrones.

Esta etapa consta de dos pasos. La Figura 2 muestra el primer paso. Este paso permite realizar una primera aproximación de los valores de confianzas de los patrones a través del pequeño conjunto preestablecido de pares hipónimo/hiperónimo que les dio origen, es decir, a partir de las semillas. Una vez determinado el valor de confianza inicial de los patrones podemos aproximar el valor de confianza inicial de los pares en el catálogo recopilado en etapas anteriores. Posteriormente, un segundo paso propaga estas estimaciones iniciales al realizar una segunda aproximación de los valores de confianza tanto de los patrones como de los pares.

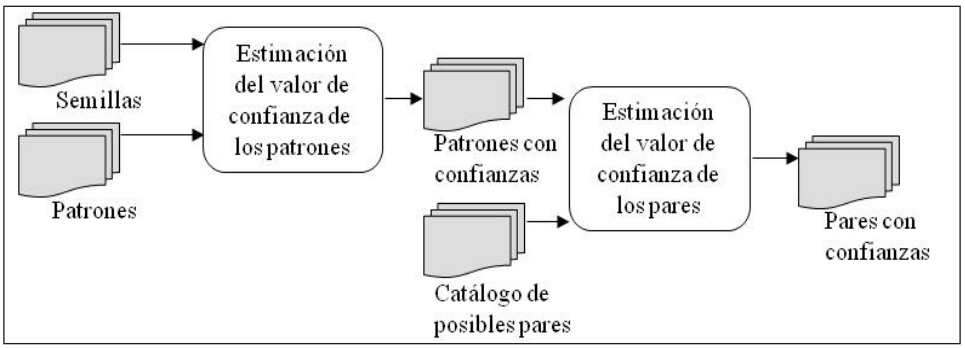


Figura 2. Primera aproximación de los valores de confianza de los pares hipónimo/hiperónimo.

La determinación del valor de confianza inicial de los patrones es una tarea importante, porque este valor ayudará a determinar el valor de confianza de los pares e impactará en la precisión final del catálogo. Básicamente, en este primer paso, tomamos en consideración el primer supuesto base, mencionado anteriormente, del proceso de estimación de los valores de confianza. El cual sugiere intuitivamente que el valor de confianza de un patrón puede determinarse mediante la cobertura y precisión del mismo.

La 'precisión' de un patrón puede ser estimada fácilmente como la porción de pares correctos extraídos por un patrón. En este mismo contexto, la 'cobertura' de un patrón puede obtenerse como el cociente del número de pares correctos extraídos por un patrón entre el número total de pares correctos en la colección. Sin embargo, la estimación de la cobertura de un patrón no es una tarea trivial, debido al desconocimiento del conjunto total de pares correctos en la colección de textos; más aún dado que los fragmentos provienen de la Web.

Dado este inconveniente, algunos trabajos han optado por medir el valor de confianza de un patrón usando la 'información mutua', estudiada ampliamente por Church y Hanks (1990). Esta medida es usada para medir el grado de asociación entre los patrones y los pares extraídos. Sin embargo, de acuerdo con los experimentos mostrados en (Ortega et al., 2007), con esta medida se pueden favorecer patrones con alta cobertura o alta precisión, pero no necesariamente a aquellos que muestran un balance entre ambos. En contraste, en el presente trabajo, para solucionar el desconocimiento de información correcta, planteamos la idea clave que sustenta nuestro esquema de estimación: suponer que aquellos pares o patrones con mayor valor de confianza son correctos.

En principio, los pares con mayor valor de confianza son las semillas, pues su elección manual asegura la confiabilidad de los mismos. Por ello, para calcular la precisión y la cobertura de los patrones, se asume que las semillas son pares correctos. Además, debe recordarse que los patrones fueron descubiertos a partir de las semillas; por lo tanto, pueden aportar información valiosa para estimar el valor de confianza de los mismos. Entonces, para iniciar el cálculo de los valores de confianzas, se tomaron los patrones obtenidos por Ortega (2007) y la lista de semillas que dieron origen a los mismos.

De esta manera, la precisión de un patrón *p*, $P_\pi(p)$ puede estimarse como el cociente del número de semillas detectadas por *p* entre el número total de pares extraídos por el patrón *p* (ver Fórmula 1). En otras palabras, deseamos determinar qué porcentaje de la información detectada por *p* es correcta.

$$(1) \quad P_\pi(p) = \frac{|\text{semillas detectadas por } p|}{|\text{pares detectados por } p|}$$

La cobertura de un patrón *p*, $R_\pi(p)$, es estimada como el porcentaje de semillas extraídas por el patrón *p* de entre todas las semillas posibles (ver Fórmula 2). En este caso, deseamos cuantificar qué tanta información puede detectar *p* respecto al total de la información disponible en la colección.

$$(2) \quad R_\pi(p) = \frac{|\text{semillas detectadas por } p|}{|\text{semillas}|}$$

Ahora bien, un patrón ideal debería tener alta cobertura y al mismo tiempo, alta precisión. Desafortunadamente, generalmente estas dos medidas están inversamente relacionadas. Por ejemplo, un patrón con alta cobertura como

"<hiperónimo> de <hipónimo> y" generalmente recupera muchos pares correctos, pero a su vez recupera muchos pares incorrectos. Es decir, tiene alta cobertura pero baja precisión. Por otro lado, un patrón que es muy específico como: "la <hipónimo> es el único <hiperónimo> natural" recupera pocos pares de entre todos los posibles, pero aquellos detectados son casi en su totalidad correctos.

Como puede observarse, ambas medidas son importantes para determinar el valor de confianza de un patrón. De ahí la necesidad de plantear la Fórmula 3, la cual integra estas dos medidas en un sola para finalmente determinar el valor de confianza $c_\pi(p)$ de un patrón *p*.

$$(3) \quad c_\pi(p) = \frac{F_\pi(p)}{\max_{\forall i \in P}\{F_\pi(i)\}}$$

Donde *P* es el conjunto de patrones y $F_\pi(p)$ está definido de la siguiente manera:

$$(4) \quad F_\pi(p) = \frac{2 \cdot P_\pi(p) \cdot R_\pi(p)}{P_\pi(p) + R_\pi(p)}$$

$F_\pi(p)$ es una adaptación de la medida *F* tradicional propuesta por Van Rijsbergen (1979), la cual se acomoda claramente a nuestro problema, ya que combina la precisión y la cobertura en una sola medida.

Una vez establecido el valor de confianza inicial de cada patrón, se procede a estimar el valor de confianza de los pares hipónimo/hiperónimo, en un proceso análogo al descrito en párrafos anteriores. Aquí, la información correcta es representada por un conjunto de patrones llamados 'patrones relevantes'. Los patrones relevantes son aquellos cuyo valor de confianza es mayor al valor de confianza promedio del conjunto total de patrones.

Con esta definición en mente y recordando el segundo supuesto del proceso de estimación, el cual establece que un par confiable debería ser extraído por un gran número de patrones cuyo valor de confianza sea relevante, se puede deducir que un par ideal debería mantener alta cobertura y a su vez, alta precisión. La estimación de estos conceptos se señala a continuación.

La precisión de un par hipónimo/hiperónimo *t*, Pσ(*t*), puede estimarse como el cociente del número de patrones relevantes que detectaron t entre el número total de patrones que detectan dicho par (incluyendo patrones no relevantes) (ver Fórmula 5).

$$(5) \quad P_\sigma(t) = \frac{|\text{patrones relevantes detectando } t|}{|\text{patrones detectando } t|}$$

La cobertura de un par *t*, Rσ(*t*), es estimada como el porcentaje de patrones relevantes que detectaron *t* entre el total de patrones relevantes (ver Fórmula 6).

$$(6) \quad R_\sigma(p) = \frac{|\text{patrones relevantes detectando } t|}{|\text{patrones relevantes}|}$$

Naturalmente, nosotros necesitamos integrar estas dos medidas en una sola. Razón por la cual se adapta nuevamente la medida *F* para medir la calidad de los pares como se muestra en la Fórmula 7.

$$(7) \quad F_\sigma(t) = \frac{2 \cdot P_\sigma(t) \cdot R_\sigma(t)}{P_\sigma(t) + R_\sigma(t)}$$

Si bien, la Fórmula 7 figura como apropiada, por sí sola no puede determinar el valor de confianza de un par; principalmente por dos situaciones. La primera, hace referencia a los valores nulos que esta medida asigna a los pares extraídos por pocos o muchos patrones, pero que no operan como relevantes. Para mayor claridad, considere los patrones mostrados en la Tabla 3 cuyo valor de confianza promedio es 0,27. Algunos de estos patrones extraen el par (cardiopatía isquémica', 'enfermedad') mostrado en el Cuadro 3. En este escenario, los patrones del Cuadro 3 no forman parte del conjunto de patrones relevantes. En consecuencia, el valor de F_σ para el par ('cardiopatía isquémica', 'enfermedad') es igual a 0, pues no existen patrones relevantes que lo extraigan. Por consiguiente, aun cuando el par es correcto, la Fórmula 7 no podría detectarlo como confiable.

Tabla 3. Lista de patrones, su valor de confianza y la indicación de relevancia.

Patrón	Cofianza	¿Relevante?
la <hipónimo> es una <hiperónimo> que	0,64	Sí
la <hipónimo> es la <hiperónimo>	0,31	Sí
de <hipónimo> o <hiperónimo>	0,18	No
la <hipónimo> como una <hiperónimo>	0,16	No
la <hipónimo> una <hiperónimo>	0,06	No

Por otro lado, para explicar la segunda situación, recuerde que el objetivo es ordenar los pares del catálogo de acuerdo con su valor de confianza. En este contexto, la sola utilización de $F_\sigma(t)$ propiciaría que dos o más pares extraídos por el mismo número de patrones tanto relevantes como no relevantes obtuvieran el mismo valor de confianza y no podríamos distinguir cuál de ellos es más confiable.

A fin de resolver estos dos inconvenientes, se sugiere considerar la 'calidad' de los patrones extrayendo un par. Por lo tanto, se propone estimar el valor de confianza de un par *t* usando la Fórmula 8. Donde |*P'*| es el número de patrones extrayendo el par *t* en cuestión.

$$(8)\quad c_\sigma(t) = \frac{(\lambda_1 F_\sigma(t) + (1-\lambda_1) v_\sigma(t))^* \mid P' \mid}{\max_{\forall q \in T} \{c_\sigma(q)\}}$$

En efecto, la Fórmula 8 conjuga los beneficios de la medida *F* con la calidad de los patrones. La calidad de los patrones v_σ para un par *t*, es decir, $v_\sigma(t)$, está determinada por la Fórmula 9. Donde *P'* representa el conjunto de patrones relevantes extrayendo el par *t*, y es un subconjunto del conjunto total de patrones.

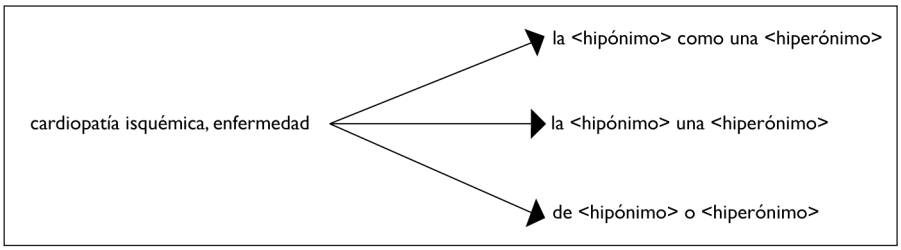
$$(9)\quad v_\sigma(t) = \sum_{p \in P'} c_\pi(p)$$

A su vez, el coeficiente λ_1 es usado para reflejar la importancia de los dos componentes de la Fórmula 8. De esta forma, se tiene la oportunidad de dar más importancia a la medida *F* o a la calidad de los patrones según la interpretación del problema. En nuestro caso se utilizó $\lambda_1 = 0{,}75$ para explotar al máximo los beneficios de la medida *F* y usar la calidad de los patrones como un factor de corrección de los valores de confianza.

Por último, en la Fórmula 8, la multiplicación por el factor |*P'*|, premia a aquellos pares extraídos por un gran número de patrones. En otras palabras, entre mayor sea el número de patrones extrayendo un par, mayor será el valor de confianza estimado.

Una vez calculado el valor de confianza de cada par en el catálogo, hemos terminado el cálculo de los valores de confianza inicial. Sin embargo, estos valores iniciales son relativos al conjunto inicial de semillas dadas. Para disminuir esta dependencia, primero, recalculamos el valor de confianza de los patrones sustituyendo el conjunto de semillas (información correcta) por un conjunto de pares 'relevantes', es decir, aquellos que tienen un valor de confianza por arriba del promedio.

En esencia, la diferencia entre el cálculo del valor de confianza inicial de un patrón y el cálculo del valor de confianza propagado del patrón, radica en que en la primera contamos con la certeza de



Cuadro 3. Patrones extrayendo el par (cardiopatía isquémica, enfermedad).

que las semillas representan información correcta. Y en la segunda, representamos está 'información correcta' con los pares de mayor valor de confianza. Por consiguiente, es necesario tomar en cuenta, además de la adaptación a la medida *F*, la 'calidad' de los pares extraídos por un patrón. De esta manera, el valor de confianza de un patrón *p*, $c_\pi(p)$, que extrae un conjunto de pares *T'*, se estima con la Fórmula 10.

$$(10)\quad c_\pi(p) = \frac{(\lambda_1 F_\sigma(p) + (1-\lambda_1) v_\sigma(p)) * \frac{1}{|T'|}}{\max_{\forall q \in P}\{c_\sigma(q)\}}$$

Donde $v_\pi(p)$ es una función que captura la calidad de un patrón *p* (ver Fórmula 11). El factor 1/|*T'*| es un valor que castiga a aquellos patrones generales que extraen muchos pares pero en su mayoría incorrectos, es decir, que tienen amplia cobertura, pero no necesariamente alta precisión.

$$(11)\quad v_\pi(p) = \sum_{t \in T'} c_\sigma(t)$$

Finalmente, el método entrega como respuesta el catálogo de pares ordenado por sus valores de confianza.

4. Resultados

El método propuesto se evaluó al buscar hipónimos para un vocabulario conformado por 5 términos: 'banco', 'enfermedad', 'felino', 'profesión' y 'roca'. Las diferentes áreas conceptuales de los términos permiten estudiar el comportamiento del método en diversos dominios de aplicación.

Para formar las consultas, los 43 patrones, mostrados previamente en la Tabla 2, fueron instanciados con los términos del vocabulario. Posteriormente, con las consultas se recopiló un conjunto significativo de fragmentos de texto (8,6 MB) provenientes de Internet. Como es de imaginar, mientras más datos se tengan, mejores serán los resultados alcanzados. Finalmente, al aplicar los patrones a los fragmentos recopilados se recuperó el catálogo de posibles pares hipónimos/hiperónimos. El catálogo quedó conformado por 851 entradas distribuidas como se exhibe en la Tabla 4.

Tras haber derivado este conjunto de pares candidatos, se pasó a realizar la primera aproximación de los valores de confianza de patrones y pares. Para determinar el valor de confianza inicial de los patrones se utilizó el conjunto de 25 semillas planteado por Ortega (2007). Una vez realizada esta primera estimación del valor de confianza de los patrones, se determinaron los patrones relevantes –aquellos patrones cuyo valor de confianza era superior al valor de confianza promedio–. Del conjunto total de 43 patrones se identificaron 10 como relevantes. La Tabla 5 muestra la lista de los patrones relevantes con sus valores de confianza correspondientes.

A partir de la Tabla 5, se muestra que el valor de confianza inicial de los patrones es una buena aproximación, pues en la lista existen patrones guardando un equilibrio entre precisión y cobertura. De hecho los patrones generales fueron localizados en las últimas posiciones. Por ejemplo, el patrón 'de' <hipónimo> 'y' <hiperónimo> es un patrón muy general de baja calidad; por ello no se clasificó como relevante; e incluso, fue colocado en la penúltima posición de la lista con un valor de confianza de 0,04. Un caso similar se presenta con el patrón <hipónimo> 'de' <hiperónimo> cuyo valor de confianza fue de 0,01 y se ubicó al final de la lista.

Tabla 4. Términos asociados a hipónimos.

Término	Hipónimos asociados
Banco	193
Enfermedad	307
Felino	9
Profesión	226
Roca	116
Total	851

Tabla 5. Listado de los patrones relevantes y sus valores de confianza.

No.	Patrón	Confianza
1	el <hipónimo> es el único <hiperónimo>	1,00
2	el uso de la <hipónimo> como <hiperónimo>	0,83
3	el <hipónimo> es uno de los <hiperónimo> más	0,57
4	de la <hipónimo> como <hiperónimo> de	0,42
5	de las <hipónimo> como <hiperónimo>	0,40
6	las <hipónimo> son una <hiperónimo>	0,31
7	el <hipónimo> es un <hiperónimo> que	0,27
8	el <hipónimo> es el <hiperónimo> que	0,26
9	de <hiperónimo> como <hipónimo> y	0,25
10	la <hipónimo> es un <hiperónimo>	0,17

Después, el valor de confianza inicial de los patrones ayudó a determinar el valor de confianza inicial de los 851 pares del catálogo. Al terminar este primer ciclo de estimaciones se propagaron los valores de confianza de los patrones, pero en esta ocasión, usando los pares relevantes. Se identificaron 20 pares como relevantes y con ellos se recalcularon los valores de confianza de los patrones y, subsecuentemente, los valores de confianza de los pares.

Después de esta segunda estimación se ordenó el catálogo de acuerdo con el valor de confianza de los pares. Aquí, para evaluar el desempeño general del método propuesto se obtuvo la precisión del catálogo. Para reportar la precisión se evaluaron manualmente los 200 primeros pares del catálogo, es decir, aquellos con mayor valor de confianza.

El Gráfico 1 muestra el porcentaje de pares correctos según se incrementa el número de pares. Como puede observarse, en las primeras 40 posiciones del catálogo ordenado tenemos más del 90% de pares correctos. Esto sucede cuando estimamos el valor de confianza inicial de los pares. Ahora bien, cuando realizamos la propagación de los valores de confianza se percibe una precisión del 100% en los 20 primeros pares manteniendo un valor de confianza por encima de la reportada en la estimación inicial en los 50 primeros pares (los más confiables). Así entonces, se observa el alza de precisión durante la propagación de estimación, con respecto a la estimación inicial, logrando una precisión de 80% para los 200 primeros pares. En la parte central de las curvas existe una variación de precisión, pero no por más de uno o dos pares correctos.

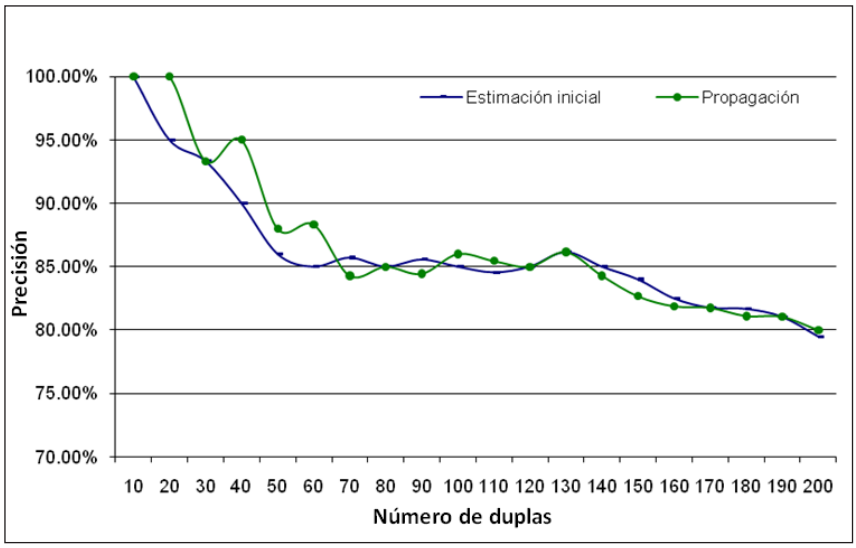


Gráfico 1. Porcentaje de pares correctos sobre el catálogo ordenado.

Tabla 6. Listado de los primeros cinco pares hipónimo/hiperónimo identificados por el sistema para los hiperónimos 'enfermedad', 'banco', 'felino', 'profesión' y 'roca', con sus respectivos valores de confianza.

Enfermedad		Banco	
obesidad 1,00	tuberculosis 1,00	BID 0,41	HSBC 0,40
cáncer 0,83	caries 0,99	cual* 0,35	BID 0,33
depresión 0,72	obesidad 0,96	BBVA 0,32	cual* 0,33
tuberculosis 0,70	gripe aviar 0,89	HSBC 0,31	Nación* 0,31
diabetes 0,70	diabetes 0,89	República* 0,28	BBVA 0,26

Felino	
jaguar 0,42	jaguar 0,55
puma 0,37	puma 0,31
lince ibérico 0,17	lince 0,16
lince 0,15	margay 0,15
margay 0,12	lince ibérico 0,17

Profesión		Roca	
medicina 0,82	política 0,94	basalto 0,26	areniscas 0,14
enfermería 0,68	enfermería 0,93	pórfidos 0,20	calizas 0,09
docencia 0,58	medicina 0,89	granito 0,16	rocas sedimentarias 0,06
psicología 0,56	psicología 0,78	mármol 0,16	basalto 0,01
política 0,45	abogacía 0,70	lava 0,16	lava 0,01

Como muestra de los resultados arrojados por el sistema, la Tabla 6 presenta un listado de los hipónimos detectados automáticamente para los cinco hiperónimos en cuestión. La primera columna de cada concepto exhibe los hipónimos con mayor valor de confianza inicial. Por su parte, la segunda columna muestra los pares que obtuvieron el mayor valor de confianza en la etapa de propagación. Para cada pareja se muestra el valor de confianza estimado por el método. A través del valor de confianza mostrado, se puede deducir que desde la estimación inicial se obtienen resultados favorables. No obstante, la propagación de los valores de confianza hace que algunos pares correctos que en principio se ubicaban en las últimas posiciones del catálogo, ahora se presenten en las primeras.

Como puede observarse, la mayoría de los pares asociados son correctos, salvo algunos casos. Durante la estimación inicial tenemos los pares incorrectos: 'banco/cual' y 'banco/República'. Posteriormente, durante la propagación de los valores de confianza tenemos los pares incorrectos: 'banco/cual' y 'banco/Nación'.

5. Discusión de los resultados

Como puede observarse en los resultados reportados la adaptación de la 'medida F' como criterio para evaluar la confianza de los patrones, así como de los pares obtenidos tuvo efectos favorables. Asimismo la estrategia de utilizar un proceso entrelazado –donde el cálculo de los valores de confianza de los patrones dependa de los valores de confianza de los pares y viceversa– demostró ser de utilidad al ver el cambio de los valores de confianza durante la primera fase de aproximación de los valores de confianza y la etapa de propagación. El caso más claro es en el cambio del valor de confianza de los patrones. La Tabla 7 muestra la lista de los 10 patrones más confiables después de la etapa de propagación. Como puede observarse, existe un claro reacomodo respecto a los valores de confianza inicial (véase la Tabla 5).

Respecto al resultado final dentro del catálogo de pares, a pesar de que el cambio en la precisión del catálogo no es notoria entre la estimación inicial y la propagación de los valores de confianza (véase el Gráfico 1), sí existe un cambio en los valores de

Tabla 7. Listado de los patrones relevantes y sus valores de confianza después de la etapa de propagación.

No.	Patrón	Confianza
20	de <hipónimo> y otras <hiperónimo>	1,00
35	de <hipónimo> y otros <hiperónimo>	0,99
32	el <hipónimo> es el único <hiperónimo>	0,96
7	el <hipónimo> es un <hiperónimo> que	0,87
37	los <hipónimo> y otros <hiperónimo>	0,75
3	el <hipónimo> es uno de los <hiperónimo> más	0,74
17	que la <hipónimo> es una <hiperónimo>	0,70
39	la <hipónimo> y otros <hiperónimo>	0,65
15	la <hipónimo> es una <hiperónimo> que	0,61
13	el <hipónimo> es un <hiperónimo> de	0,61

confianza calculados para cada par. Gracias a ello es posible determinar más fácilmente un umbral del valor de confianza que permita reunir un mayor número de pares correctos.

Por otro lado, el método fue probado en la extracción de pares de hipónimos/hiperónimos para 'enfermedad', 'banco', 'felino', 'profesión' y 'roca'. Tal como se menciona en párrafos anteriores se obtuvieron 851 entradas para este conjunto de términos. Donde los términos 'enfermedad' y 'profesión' obtuvieron 307 y 226 entradas respectivamente (véase la Tabla 4). Esta cantidad de información permitió determinar con mejor confianza los pares hipónimo/hiperónimo asociados, incluso puede verse un incremento en los valores de confianza entre la primera aproximación después de la propagación (veáse la Tabla 6). El caso de los términos 'roca' y 'banco' se obtienen menos entradas y aunado a la ambigüedad de estos términos dificulta la correcta identificación automática de pares hipónimo/hiperónimo confiables. Por último, a pesar de que se identificaron pares correctos para el término 'felino', los valores de confianza asociados a estos pares son un reflejo de las pocas entradas que cuenta el catálogo asociadas a este término. En general, es posible observar que el método se comportará mejor mientras más entradas existan en el catálogo de posibles pares.

CONCLUSIONES

En este trabajo se ha expuesto un experimento, desde el proceso de desarrollo hasta los resultados obtenidos, orientado a la identificación de pares hipónimo/hiperónimo en documentos situados en Internet, tomando en cuenta una serie de patrones léxicos. El método utilizado se sustenta en dos supuestos sencillos que son capturados a través de los conceptos de precisión y cobertura. Así un patrón es más confiable mientras mayor sea la cantidad de pares confiables detectados por el mismo; y un par es más confiable mientras mayor sea la cantidad de patrones que lo detectan. Los resultados muestran la factibilidad del esquema de extracción de pares hipónimo/hiperónimo.

Este método es un claro ejemplo del uso de métodos híbridos para tareas de extracción de información, que combina conocimientos lingüísticos, computacionales y estadísticos, en aras de concretar resultados confiables.

Por otra parte, respecto a los patrones léxicos considerados, cabe señalar que si bien el listado aquí propuesto no agota todas las opciones que tienen los hipónimos e hiperónimos de ser expresados en textos, el hecho de que permitan alcanzar niveles considerables de precisión, hace que tales patrones puedan ser vistos como pertinentes para identificar automáticamente palabras que refieran a una relación de hiponimia/hiperonimia.

Tomando en cuenta tal pertinencia, una proyección de este trabajo sería evaluar si dichos patrones, así como el sistema implementado para este experimento, son capaces de reconocer hipónimos e hiperónimos en otras lenguas. Haciendo los ajustes necesarios, particularmente atendiendo al tipo de patrones léxico-sintácticos que sigan otros idiomas para codificar relaciones de hiponimia/hiperonimia, podría considerarse sustentable ampliar el campo de aplicación de los métodos y las herramientas generadas en este experimento.

REFERENCIAS BIBLIOGRÁFICAS

Águila, G. (2006). Las nuevas tecnologías al servicio de la lexicografía: Los diccionarios electrónicos. En M. Villayandre (Ed.), *Actas del XXXV Simposio Internacional de la Sociedad Española de Lingüística* (pp. 1-23). León: Universidad de León.

Ahonen-Myka, H. (2002). Discovery of frequent word sequences in text source. *En Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery.* London: U. K.

Barbu, V. (2008). Hyponymy patterns: Semi-automatic extraction, evaluation and inter-lingual comparison. En P. Sojka, A. Horak, I. Kopecek & P. Karel (Eds.), *Text, Speech and Dialogue* (pp. 37-44). Berlin: Springer.

Baroni, M. & Bisi, S. (2004). Using cooccurrence statistics and the Web to discover synonyms in a technical language. En *Proceedings of the 4th International Conference on Language Resources and Evaluation.* Lisbon: ELDA.

Blohm, S., Cimiano, P. & Stemle, E. (2007). Harvesting relations from the Web: Quantifiying the impact of filtering functions. En *Proceedings of the 22nd National Conference on Artificial intelligence*. Vancouver: AAAI Press.

Buitelaar, P., Cimiano, P. & Magnini, B. (2007). *Ontology learning from text: Methods, evaluation and applications*. Amsterdam: IOS Press.

Cabré, T., Estopà, R. & Vivaldi, J. (2001). Automatic term detection. En D. Bourigault, C. Jaquemin & M. C. L'Homme (Eds.), *Recent Advances in Computational Terminology* (pp. 53-87). Amsterdam: John Benjamins.

Church, K. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics, 16*(1), 22-29.

Cimiano, P. (2006). *Ontology learning and population from text, algorithms, evaluation and applications*. Nueva York: Springer.

Cruse, D. (1986). *Lexical semantics.* Cambridge: Cambridge University Press.

Denicia, C., Montes, M., Villaseñor, L. & García, R. (2006). A text mining approach for definition question answering. En *Proceedings of FinTAL*. Berlin: Springer.

Fellbaum, C. (Ed.) (1998). *WordNet: An electronic lexical database*. Cambridge: MIT Press.

García Hernández, R., Martínez-Trinidad, F. & Carrasco-Ochoa, A. (2006). A new algorithm for fast discovery of maximal sequential patterns in a document collection. En *Proceedings of International Conference on Computational Linguistics and text Processing*. Mexico City: Mexico.

Gelbukh, A. & Sidorov, G. (2006). *Procesamiento automático del Español con enfoque en recursos léxicos grandes*. México: Instituto Politécnico Nacional.

Girju, R., Badulescu, A. & Moldovan, D. (2006). Automatic discovery of part–whole relations. *Computational Linguistics, 32*(1), 83-135.

Hearts, M. (1992). Automatic acquisition of hyponyms from large text corpora. En P*roceedings of Conference COLING*. Nantes: Association for Computational Linguistics.

Llisterri, J. (2003). Lingüística y tecnologías del lenguaje. Lynx. *Panorámica de estudios lingüísticos*, 2, 9-71.

Lucero, C., Pinto, D. & Jiménez, H. (2004). *A tool for automatic detection of antonymy relations*. Ponencia presentada en el IX Ibero-American Conference on Artificial Intelligence, Puebla, México.

Miller, G. (1998). Nous in WordNet. En C. Fellbaum. (Ed.), *WordNet: An electronic lexical database* (pp. 23-46). Cambridge: MIT Press.

Ortega, R. (2007). *Descubrimiento automático de hipónimos a partir de texto no estructurado*. Tesis de maestría en Ciencias Computacionales, Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, México.

Ortega, R., Villaseñor, L. & Montes, M. (2007). Using lexical patterns for extracting hyponyms from the Web. En *Proceedings of MICAI.* Berlin: Springer.

Pasca, M. (2004). Acquisition of categorized named entities for Web search. En *Proceedings of the 13th ACM international conference on Information and knowledge management*. Washington: ACM.

Pantel, P. & Pennacchiotti, M. (2006). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. En *Proceedings of Conference on Computational Linguistics Association for Computational Linguistics.* Sydney: ACL.

Pennacchiotti, M. & Pantel, P. (2009). Entity extraction via ensemble semantics. En *Proceedings of Conference on Empirical Methods in Natural Language Processing*. Singapore: ACL.

Rojo, G. (2008). *Lingüística de corpus y lingüística del Español*. Conferencia Magistral presentada en el XV Congreso Internacional ALFAL, Montevideo, Uruguay.

Sierra, G., Alarcón, R., Aguilar, C. & Bach, C. (2008). Definitional verbal patterns for semantic relation extraction. *Terminology, 14*(1), 74-98.

Snow, R., Jurafsky, D. & Ng, A. (2006). Semantic taxonomy induction from heterogeneous evidence. En *Proceedings of the 21st International Conference on Computational Linguistics*. Sydney: Association for Computational Linguistics.

Van Rijsbergen, C. (1979). *Information retrieval.* Ontario: Butterworths.

Wilks, Y., Slator, B. & Guthrie, L. (1996). *Electric words*. Cambridge: MIT Press.

* Los autores desean expresar su agradecimiento a los miembros del Laboratorio de Tecnologías del Lenguaje del Instituto Nacional de Astrofísica, Óptica y Electrónica y a los miembros del Grupo de Ingeniería Lingüística de la Universidad Autónoma de México por el apoyo brindado para la realización de este trabajo. De igual forma agradecen el aporte financiero otorgado por el Consejo Nacional de Ciencia y Tecnología de México a través de la beca de posgrado 223498 y los proyectos 82050, 106013 y 134186.