Desambiguación y expansión semántica automáticas de la consulta utilizando WordNet

Automatic semantic disambiguation and query expansion using WordNet.

Cristina Bender, Marcos Belén, Claudia Deco
Facultad de Ciencias Exactas, Ingeniería y Agrimensura
Universidad Nacional de Rosario
bender@fceia.unr.edu.ar, marcos.e.belen@gmail.com, deco@fceia.unr.edu.ar

Abstract

Most users seeking information on the web do not know how to build an appropriate search strategy. For this, they have difficulty finding useful information efficiently, and many of the retrieved documents have little or nothing to do with his/her search interest. In addition, many relevant documents are not found by search engines on the Web since natural language is used. This paper describes a method to expand the user's query automatically using linguistic resources, in order to increase the chances of retrieving the desired information. In particular, we propose the automatic disambiguation of terms and the automatic semantic expansion using the WordNet resource.

Keywords: Semantic expansion, Automatic disambiguation, Linguistic resources, WordNet

Resumen

La mayoría de los usuarios que buscan información en la web no saben construir una estrategia de búsqueda adecuada. Por esto, tienen dificultades para encontrar información valiosa de manera eficiente, y muchos de los documentos devueltos tienen poco o nada que ver con su interés de búsqueda. Además, muchos documentos relevantes no son encontrados por los motores de búsqueda dado que en la Web se utiliza lenguaje natural. En este trabajo se describe un método que permite expandir la consulta del usuario de forma automática mediante recursos lingüísticos, buscando incrementar las posibilidades de recuperar la información deseada. En particular, se propone la desambiguación automática de términos y la expansión semántica automática, utilizando el recurso WordNet

Palabras clave: Expansión semántica, Desambiguación automática, Recursos lingüísticos, WordNet

1. INTRODUCCIÓN

La Web se ha convertido en un repositorio universal de cultura y conocimiento humano el cual ha permitido compartir de una manera sin precedentes la información. Sin embargo, encontrar en la Web información útil es frecuentemente una tarea difícil y tediosa. El objetivo principal de los Sistemas de Recuperación de Información (RI) es encontrar información útil o relevante para el usuario. Para ser efectivos en este intento de satisfacer las necesidades de información del usuario, el sistema de RI debe saber *cómo* interpretar el contenido de los documentos y ordenarlos de acuerdo a un grado de relevancia respecto a la consulta. Esta interpretación del contenido implica extraer información sintáctica y semántica del texto del documento y usarla para recuperar información relevante. La dificultad no es sólo saber cómo extraer esta información sino también como usarla para decidir relevancia. Para evaluar los resultados de una consulta, se utilizan dos indicadores: precisión (que mide el ratio entre la cantidad de documentos relevantes recuperados y la cantidad de documentos recuperados) y recall (que mide el ratio entre la cantidad de documentos relevantes recuperados y la cantidad total de documentos relevantes de la colección).

Un problema de la RI es que usualmente trata con texto en lenguaje natural el cual no siempre está bien estructurado y muchas veces es semánticamente ambiguo. Los sistemas de recuperación que emplean técnicas de indexación automática para crear términos representativos de documentos escritos en lenguaje natural, deben tratar con problemas de polisemia y sinonimia. La polisemia (palabras con más de un significado) degrada la precisión causando falsas coincidencias, mientras que la sinonimia (múltiples palabras teniendo el mismo significado) degrada la proporción de documentos relevantes que son recuperados, causando la pérdida de verdaderas coincidencias conceptuales. En principio, la polisemia y la sinonimia se pueden manejar o controlar asignando a diferentes sentidos de una palabra diferentes identificadores del concepto, y asignando el mismo identificador de concepto a los sinónimos. En la práctica esto requiere procedimientos que sean capaces de reconocer sinónimos, y que puedan no sólo detectar los usos de diferentes sentidos de una palabra sino que también puedan resolver qué significado se desea en cada caso. Un elemento a tener en cuenta es la utilización de recursos lingüísticos, tales como tesauros y diccionarios. Un recurso lingüístico muy utilizado es WordNet ([1], [2]).

Para buscar información, el usuario describe su interés mediante una consulta utilizando términos o frases. En este trabajo se presenta un método automático para construir una estrategia de búsqueda que permita ayudar al usuario a encontrar información útil de manera eficiente. Para esto, por un lado se amplía la consulta con sinónimos definidos en WordNet para lidiar con el problema de la sinonimia; y por otro, se busca el sentido correcto de cada sustantivo o frase hallada en la consulta mediante un método de desambiguación que se basa en las relaciones existentes entre los términos utilizando el contenido semántico de WordNet. Esta estrategia de búsqueda servirá de base para la búsqueda en la Web en general, y en cualquier sistema de recuperación de información.

El resto del trabajo se organiza de la siguiente forma: en la Sección 2 se describen algunos recursos lingüísticos. En la Sección 3 se presentan trabajos relacionados. En la Sección 4 se describe el método propuesto. Luego se presenta la propuesta de un prototipo y casos de uso. Finalmente se presentan las conclusiones.

2. Recursos Lingüísticos

WordNet es un sistema léxico construido manualmente por George Miller y sus colegas en el laboratorio de Ciencias Cognitivas de la Universidad de Princeton ([1], [2]). El objetivo inicial era construir un diccionario en el que se pudiera buscar de manera conceptual en lugar de sólo

alfabéticamente. El objeto básico en WordNet es un conjunto de sinónimos exactos llamados synsets. Por definición, cada synset es un sentido o significado diferente de cada palabra, es decir, cada synset representa un concepto. WordNet tiene cuatro divisiones principales: sustantivos, verbos, adjetivos y adverbios. Dentro de una división, los synsets están organizados por las relaciones léxicas definidas sobre ellos. Para los sustantivos, la única división utilizada en este trabajo, las relaciones léxicas incluidas son: antonimia, hiperonimia/hiponimia, y tres diferentes relaciones de meronimia/holonimia. La versión 3.0 contiene 117798 sustantivos, organizados en 82115 synsets.

En este trabajo, WordNet se utilizó para la desambiguación de los términos que componen la consulta y posteriormente la búsqueda de los sinónimos y homónimos para la generación de la estrategia de búsqueda. Para esto, se ha realizado una conversión de las relaciones de sinonimia, hiperonimia e hiponimia de la red semántica de WordNet al formato XML. En la Figura 1 se puede observar parte del archivo.

```
- <synset>
    <id>02139199</id>
    <num_words>02</num_words>
  - <synonyms>
      <term>bat</term>
      <term>chiropteran</term>
    </synonyms>
  - <hyperonyms>
      <term>01886756</term>
    </hyperonyms>
  - <hyponyms>
      <term>02139671</term>
      <term>02141306</term>
    </hyponyms>
    <descripcion>nocturnal mouselike mammal with forelimbs modified to form
      membranous wings and anatomical adaptations
      for echolocation by which they navigate</descripcion>
  </synset>
```

Figura 1. Parte de la base de datos de WordNet: Synset correspondiente al término bat.

Dentro del archivo los synsets se distinguen por la etiqueta <synset>. La etiqueta <id> contiene el identificador del synset, <num_words> indica la cantidad de términos que componen el synset, <synonyms> contiene los términos sinónimos cada uno identificado con una etiqueta <term>, <hyperonyms> contiene los identificadores de los synsets “padres” (términos más generales), en <hyponyms> se encuentran los identificadores de los synsets “hijos” (términos más específicos), y en <descripcion> se describe el concepto que representa ese synset.

WordNet mantiene en otro archivo la asociación de cada término con sus respectivos synsets. Este archivo permite encontrar los distintos sentidos de una palabra o frase, por lo cual también ha sido convertido a XML y se puede ver un ejemplo en la Figura 2.

```
- <noun>
    <id>bat</id>
    <synset>02139199</synset>
    <synset>00458456</synset>
    <synset>04292414</synset>
    <synset>03132076</synset>
    <synset>02806379</synset>
  </noun>
- <noun>
    <id>bat_boy</id>
    <synset>09843443</synset>
  </noun>
- <noun>
    <id>bat_mitzvah</id>
    <synset>07454196</synset>
  </noun>
```

Figura 2. XML con términos y synsets a los que corresponde.

En esta figura, se resalta en amarillo el identificador del synset descripto en la Figura 1. Cada sustantivo junto con todos los synsets que tenga asociado se encuentran entre las etiquetas <noun></noun>. Dicho sustantivo se distingue por la etiqueta <id> y cada uno de los sentidos del mismo es identificado por un par <synset></synset>.

3. Trabajos Relacionados

En [3] se realizan experimentos de recuperación de documentos usando conocimiento semántico. En un primer conjunto de experimentos, se utilizan sinónimos e hipónimos obtenidos de WordNet para enriquecer las consultas. En la segunda parte de la experimentación se utiliza un sistema de desambiguación (Word Sense Disambiguation - WSD) para tratar el problema de la polisemia. También experimentan con tesauros especializados. Los resultados conducen a que el uso de sinónimos no necesariamente decrece la precisión, que el uso de WSD no necesariamente decrece el recall y que el uso de recursos especializados puede ser útil para mejorar el rendimiento.

En [4] se presenta un método para WSD automático de sustantivos dentro de un conjunto de sustantivos relacionados. La esencia del algoritmo de desambiguación es el cálculo de lo que llama similitud semántica usando la taxonomía de WordNet. Esta similitud se basa en la observación de que cuando dos palabras polisémicas son similares, sus synsets ancestros proveen información sobre cuál es el sentido más relevante de cada palabra. Remarcan también la diferencia entre similitud y relación entre dos palabras: similitud es una relación más especializada que relación o asociación, por ejemplo médico y enfermedad están altamente relacionadas pero no son similares.

[5] analizan la integración de la tecnología disponible para el tratamiento del lenguaje natural en el desarrollo de un metabuscador que alcance un mayor grado de acierto en la recuperación de información así como en el tratamiento posterior de los documentos recuperados. En particular, describen el proceso realizado para la extensión de las consultas de los usuarios mediante información lingüística empleando dos recursos léxicos para el castellano: ARIES para el tratamiento de la morfología y EuroWordNet para el tratamiento de la semántica. Sin embargo, plantean que este último recurso presenta algunos problemas para la RI dado que no incluye relaciones semánticas con referencias cruzadas entre categorías gramaticales y las distinciones semánticas son muy sutiles en algunos casos (el grado de granularidad es bajo).

[6] describe un procedimiento de indexación automático que utiliza las relaciones jerárquicas de WordNet junto con un conjunto de sustantivos contenidos en un texto para seleccionar un sentido para cada sustantivo polisémico en el texto. Los resultados no fueron buenos debido a la dificultad del método utilizado para desambiguar el sentido correcto de las palabras, principalmente en consultas cortas; con poco contexto el procedimiento falla en la desambiguación.

En [7] se describe un sistema que enfoca dos problemas: la traducción de una pregunta o sentencia en lenguaje natural a una consulta usando WordNet; y la extracción de párrafos con información relevante contenidos en los documentos recuperados por motores de búsqueda. Con el primer paso se busca mejorar el recall y con el segundo mejorar la precisión.

4. Propuesta

En este trabajo se propone realizar la expansión semántica de la consulta ingresada por el usuario utilizando sinónimos y homónimos de la base de datos léxica WordNet. El proceso de expansión se presenta en la Figura 3.

La expansión se realiza de forma automática en base a un método de desambiguación de términos descripto más adelante (Figura 4). Esta expansión permitirá mejorar la precisión así como también aumentar la cantidad de documentos relevantes recuperados, al incluir términos conceptualmente equivalentes y jerárquicamente relacionados. Los módulos que realizan la expansión conceptual de una consulta ingresada por el usuario en lenguaje natural se describen a continuación.

Figura 3. Proceso de expansión semántica de una consulta de usuario.

Dada una consulta $Q=[t_1,t_2,...,t_n]$, donde t_i es uno de los términos escritos por el usuario, el módulo 1 verifica ortográficamente cada uno de ellos. La salida es un conjunto $Q'=[t'_1,t'_2,...,t'_n]$ donde cada elemento t'_i es la corrección ortográfica de su contraparte en Q.

El segundo módulo se encarga del etiquetado morfológico de los términos. Acá se detallan los rasgos de género y número de cada t'_i. Para esto se utiliza un etiquetador sintáctico (Part Of Speech tagger) tal como el de Eric Brill [8].

La consulta resultante del etiquetado es procesada en el tercer módulo, el cual está encargado del reconocimiento de nombres propios, frases y términos no significativos (stopwords).

El siguiente módulo se ocupa de hallar el sentido correcto de las palabras en base a WordNet. Si $[w_1,w_2,w_3,..,w_k]$ es la lista de sustantivos resultante del módulo anterior, para encontrar el sentido correcto de cada w_i se efectúan los siguientes pasos:

1. Si w_j o uno de sus sinónimos se encuentra en la definición de un synset de w_i, digamos S, entonces S será el sentido de w_i.
2. La definición de cada synset de w_i y w_j son comparados entre sí. La combinación que tenga el número máximo de palabras contenidas (no stopwords) en común será el sentido para w_i y también para w_j.
3. Si w_j o uno de sus sinónimos aparece en la definición de un synset S conteniendo un hipónimo de w_i, entonces el sentido de w_i será el synset S_1 el cual contiene a w_i y tiene al descendiente S.
4. Sea w_i en un synset S_1. S_1 tiene un synset hipónimo U que contiene un término h. Si h aparece en la definición de un synset S_2 que contiene a w_j entonces el sentido de w_i será S_1, y S_2 será el sentido de w_j.
5. En otro caso, el algoritmo no asigna un sentido al término y se deja al usuario esta decisión.

En la Figura 4 se observa gráficamente este proceso de desambiguación de w_i, donde w_j es elegido

en base a su proximidad con w_i en la consulta suponiendo que mientras más relacionados estén dos términos entonces más cerca estarán uno de otro en la consulta. La Figura 5 ilustra esta idea.

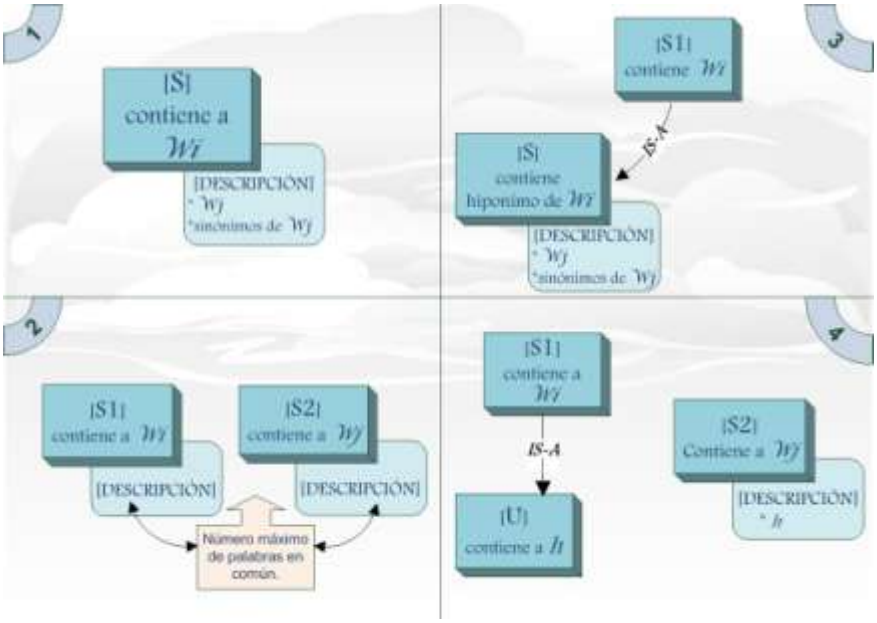

Figura 4. Proceso de desambiguación de un término w_i.

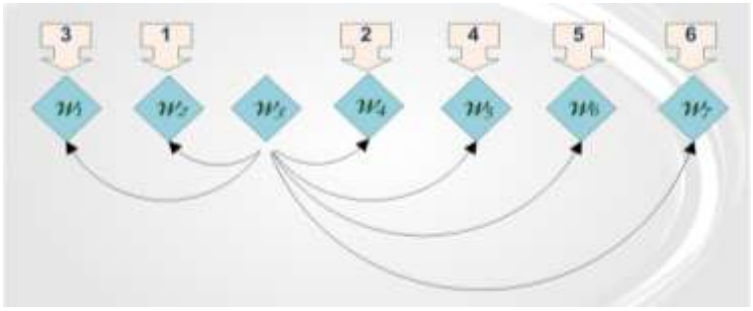

Figura 5. Orden de selección de wj para hallar el sentido de w3.

El quinto módulo (Expansión semántica) tiene como objetivo expandir la consulta original para la recuperación de documentos que son relevantes pero que no contienen los términos exactos ingresados por el usuario. Esto es, documentos que contienen términos conceptualmente equivalentes a los originales.

El módulo Generación de la estrategia produce la estrategia de búsqueda en XML. Esta estrategia es trasladada a la sintaxis del motor de búsqueda elegido por medio del séptimo módulo (Traducción).

Para ejemplificar este proceso consideremos la consulta "*cancer disease*". Estos términos son ingresados por el usuario, y se los etiqueta: [*cancer* NN], [*disease* NN]. El análisis de la existencia de frases, arroja que esta consulta no las tiene.

Para la desambiguación, en primer lugar se accede a WordNet con cada término para encontrar cuáles son los synsets a los que pertenece cada uno. Para el primer término [*cancer*], la búsqueda da como resultado 5 synsets encontrados; esto es, este término tiene 5 acepciones:

S11: **<term>**Cancer**</term>**
<term>genus_Cancer**</term>**
<descripcion>
type genus of the family Cancridae
</descripcion>
S12: **<term>**Cancer**</term>**
<term>Cancer_the_Crab**</term>**
<term>Crab**</term>**
<descripcion>
the fourth sign of the zodiac; the sun is in this sign
from about June 21 to July 22
</descripcion>
S13: **<term>**Cancer**</term>**
<descripcion>
a small zodiacal constellation in
the northern hemisphere; between Leo and Gemini
</descripcion>
S14**:** **<term>**Cancer**</term>**
<term>Crab**</term>**
<descripcion>
(astrology) a person who is born while the sun is in Cancer
</descripcion>
S15: **<term>**cancer**</term>**
<term>malignant_neoplastic_disease**</term>**
<descripcion>
any malignant growth or tumor caused by abnormal and uncontrolled cell division; it may spread to other parts of the body through the lymphatic system or the blood stream
</descripcion>

Para el término [*disease*], la búsqueda de synsets arroja 1 synset encontrado:

S21: **<term>**disease**</term>**
<descripcion>
an impairment of health or a condition of abnormal functioning
</descripcion>

En segundo lugar, se verifica si el término *disease* se encuentra entre las palabras de las descripciones (<descripcion>) de cada acepción de *cancer*. En este ejemplo, el término *disease* no se encuentra en ninguna de las 5 descripciones de *cancer*.

Por esto, se comparan los términos de la descripción de cada acepción de *cancer* con los términos de la descripción de *disease*, para ver si tienen palabras en común. Como resultado de esta comparación, se encuentra que entre la descripción de la acepción S15 de *cancer* (any malignant growth or tumor caused by **abnormal** and uncontrolled cell division; it may spread to other parts of the body through the lymphatic system or the blood stream) y la descripción de *disease* (an impairment of health or a condition of **abnormal** functioning) se tiene una palabra en común. Por lo

tanto S15 es el synset elegido para *cancer*. Este proceso de desambiguación de acepción no es necesario para *disease*, dado que pertenece a un único synset (S21).

Una vez determinadas las acepciones de cada término, el paso siguiente es encontrar los sinónimos. Estos se recuperan del campo <term> de cada synset. Para la consulta de ejemplo, estos son:

- Sinónimos en S15: <term>malignant_neoplastic_disease</term>
- Sinónimos en S21: este término no tiene sinónimos

Finalmente, se genera la consulta, resultando:

(“cancer” OR “malignant neoplastic disease”) AND (“disease”)

y se la convierte a la sintaxis del motor de búsqueda elegido. Por ejemplo, si se elige Google, la sintaxis coincide con la consulta expresada antes.

Si se utilizan los términos obtenidos en el proceso descripto, para el procesamiento de los documentos recuperados, es posible filtrarlos y así mejorar la precisión de la búsqueda. Por ejemplo, en una búsqueda en la Web, se puede realizar la comparación de los términos (incluyendo sinónimos, hipónimos, etc.) de la consulta contra los obtenidos y procesados de los metadatos (i.e. tag <keywords>) y otros rótulos (i.e. tag <title>) del HTML de cada página recuperada. Este pos-procesamiento permite ordenar los documentos de manera que los que sean más relevantes aparezcan en los primeros lugares del resultado, y aquellos documentos que no hayan sido recuperados mediante los términos exactos de la consulta serán posicionados en los últimos lugares.

5. Experimentación

Para la experimentación se desarrolló un prototipo utilizando Visual Studio 2008 .NET Framework 3.5. de Microsoft. El contenido de WordNet fue representado mediante un archivo de formato XML. Se realizaron 12 consultas sobre el repositorio de recursos educativos Ariadne [9]. De cada consulta se registró la cantidad de documentos recuperados sin expansión (aquellas que contienen sólo los términos o frases ingresados por el usuario), la cantidad de documentos recuperados con expansión (con el agregado de sinónimos e hiperónimos de las palabras claves), y la cantidad de documentos recuperados luego del proceso de filtrado y ordenado (con la comparación de los términos y frases de la consulta con los metadatos de cada documento obtenido). Además, el usuario evaluó la relevancia de cada documento respecto a su interés de búsqueda. Este registro permite evaluar la precisión de los resultados. En todos los casos, se han tomado los primeros 20 resultados.

Como caso de estudio se discute a continuación distintas alternativas de la expansión y los resultados obtenidos para la consulta del usuario “*mathematics* AND *calculus*”.

1. Consulta SIN expansión:

mathematics AND calculus

Se recuperaron 78 objetos que contienen ambas palabras claves. Algunos de los ítems recuperados tienen que ver con archivos de audio o lecciones de cálculo matemático en formato de video. En cambio otros documentos tienen una relación más "distante" al cálculo matemático, por ejemplo “8.022 Physics II: Electricity and Magnetism, Fall 2004 (MIT)” que es un curso online del MIT sobre electricidad y magnetismo y que requiere conocimientos de cálculo vectorial entre otros.

2. Consulta CON expansión (sinónimos):

("mathematics" OR "math" OR "maths") AND ("calculus" OR "infinitesimal calculus")

Se recuperaron 79 objetos. La expansión consistió en incorporar sinónimos. El uso de los sinónimos "*math*" y "*maths*" permitieron recuperar un objeto más que había sido ignorado en la búsqueda anterior. El objeto es un curso online del MIT sobre física que trata temas como fuerza, gravedad, energía, etc. dirigido a estudiantes con conocimientos de cálculo.

3. Consulta CON expansión (hiperónimos – variante 1):

("mathematics" OR "math" OR "maths") AND ("science" OR "scientific discipline")

AND ("calculus" OR "infinitesimal calculus") AND ("pure mathematics")

Esta tercera búsqueda no logró ningún resultado. Los hiperónimos, añadidos como una conjunción de disyunciones, hacen que la consulta sea más restrictiva. Si los hiperónimos son términos poco comunes existe la posibilidad de que una consulta de esta forma no recupere ningún elemento, como sucede en este caso.

4. Consulta CON expansión (hiperónimos – variante 2):

("mathematics" OR "math" OR "maths" OR "science" OR "scientific discipline")

AND ("calculus" OR "infinitesimal calculus" OR "pure mathematics")

Esta consulta recuperó 96 objetos. En la misma se utilizan los hiperónimos desde otra perspectiva: se los agrega mediante una disyunción junto con los sinónimos con los cuales están relacionados. Así, se genera una consulta más permisiva donde los resultados están relacionados con la consulta original pero abarcando temas más generales. Por ejemplo, se pueden observar resultados en los cuales se aplica el cálculo pero en otras áreas de la ciencia tales como la física o la química.

5. Consulta CON expansión (hiperónimos – variante 3):

("mathematics" OR "math" OR "maths") AND ("science" OR "scientific discipline")

AND ("calculus" OR "infinitesimal calculus")

En esta consulta se eliminó el componente más restrictivo de la variante 1, el cual provocaba que no se recuperaran ítems. La cantidad de documentos recuperados fue de 17, es decir, hubo 61 objetos que no fueron devueltos porque no cumplieron con la restricción ("*science*" OR "*scientific discipline*"). Sin embargo hubo elementos con una relevancia importante que no fueron recuperados. Tal es el caso de los siguientes objetos "*Calculus Podcasts*", "*Sequences and limits*" o "*Teach yourself limits*". Y también hubo objetos no relevantes que sí fueron devueltos, como por ejemplo "14.12 Economic Applications of Game Theory, Fall 2005 (MIT)".

Se puede concluir que la consulta expandida utilizando sinónimos es la más apropiada, ya que recupera más elementos que la consulta sin expansión y no presenta el problema de hiperónimos restrictivos. La variante 2 del uso de hiperónimos no es de utilidad ya que recupera muchos objetos no relevantes lo cual provoca una disminución de la precisión. La variante 3, supone saber cuál es el término hiperónimo que restringe una consulta de variante 1, lo que no es simple de automatizar.

Por esto, se decide realizar la expansión utilizando sólo sinónimos, dado que recupera más elementos relevantes que la consulta sin expansión, y no tiene las limitaciones del agregado de hiperónimos.

Las consultas realizadas se listan en la Tabla 1, donde se presenta para cada consulta ingresada por el usuario la correspondiente consulta expandida.

En la Tabla 2 se presentan los resultados de las consultas realizadas. Dicha tabla muestra para cada consulta los valores correspondientes a la cantidad de documentos recuperados sin expansión, con expansión y posprocesados, la cantidad de relevantes en cada caso, y la precisión. La última columna contiene los valores promedio de precisión para las 12 consultas en cada caso (sin expansión, con expansión y con filtrado).

Un problema que puede presentarse es que la expansión recupere objetos indeseados, con lo cual disminuye la precisión. Una solución es el posprocesamiento propuesto de los resultados con la realización del filtrado donde se pueden descartar aquellos objetos que no verifican que contienen en algún lugar algún término de la consulta original.

Tabla 1: Consultas realizadas

Nro consulta	**Consulta ingresada por el usuario**	**Consulta expandida**
1	plate tectonics	(“plate tectonics” OR “plate tectonic theory”)
2	encription AND algorithm	(“encriptyon” OR “encoding”) AND (“algorithm” OR “algorithmic rule” OR “algorithmic program”)
3	“computer science” AND “game theory”	(“computer science” OR “computing”) AND (“game theory” OR “theory of games”)
4	cancer AND disease AND treatment	(“cancer” OR “malignant neoplastic disease”) AND (“disease”) AND (“treatment” OR “medical care” OR medical aid”)
5	“communications protocol” AND tcp	(“communications protocol” OR “protocol”) AND (“tcp” OR “transmission control protocol”)
6	biology AND evolution	(“biology” OR “biological science”) AND (“evolution” OR “organic evolution” OR “phylogeny” OR “phylogenesis”)
7	biology AND ecology	(“biology” OR “biological science”) AND (“ecology” OR “bionomics” OR “environmental science”)
8	“artificial intelligence” AND programming	(“artificial intelligence” OR “AI”) AND (“programming” OR “programing” OR “computer programming” OR computer programing”)
9	mathematics AND calculus	(“mathematics” OR “math” OR “maths”) AND (“calculus” OR “infinitesimal calculus”)
10	“operating system” AND UNIX	(“operating system” OR “OS”) AND (“UNIX” OR “UNIX system” OR UNIX operating system”)

Tabla 2: Resultados de las consultas.

Nro Consulta	1	2	3	4	5	6	7	8	9	10	
Recuperados SIN expansión	10	2	2	9	2	91	250	53	78	37	
Recuperados CON expansion	10	22	7	9	12	94	255	80	79	37	
Recuperados Filtrados	10	10	3	9	8	35	35	34	37	26	
Relevantes SIN expansión	8	1	1	9	1	14	9	16	20	10	
Relevantes CON expansion	8	1	3	9	5	13	11	18	18	10	**Valores Promedios de Precisión**
Relevantes filtrados	8	1	2	9	4	16	13	15	18	13	
Precisión SIN expansion	0,80	0,50	0,50	1,00	0,50	0,15	0,04	0,30	0,26	0,27	0,43
Precisión CON expansion	0,80	0,05	0,43	1,00	0,42	0,14	0,04	0,23	0,23	0,27	0,36
Precisión filtrados	0,80	0,10	0,67	1,00	0,50	0,46	0,37	0,44	0,49	0,50	0,53

En general, el posprocesamiento de los documentos resultantes, mediante el filtrado y ordenado, provocó que varios elementos relevantes fueran posicionados en los primeros lugares. Por ejemplo, en la consulta 8, el documento "Logic Programming" pasó de la posición 51 a la 5, el documento "Concurrent Prolog: A Progress Report" pasó de la posición 21 a la 12, y el documento "Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project" pasó de la posición 72 a la 17.

La Figura 6 muestra gráficamente la comparación los valores de precisión de cada consulta.

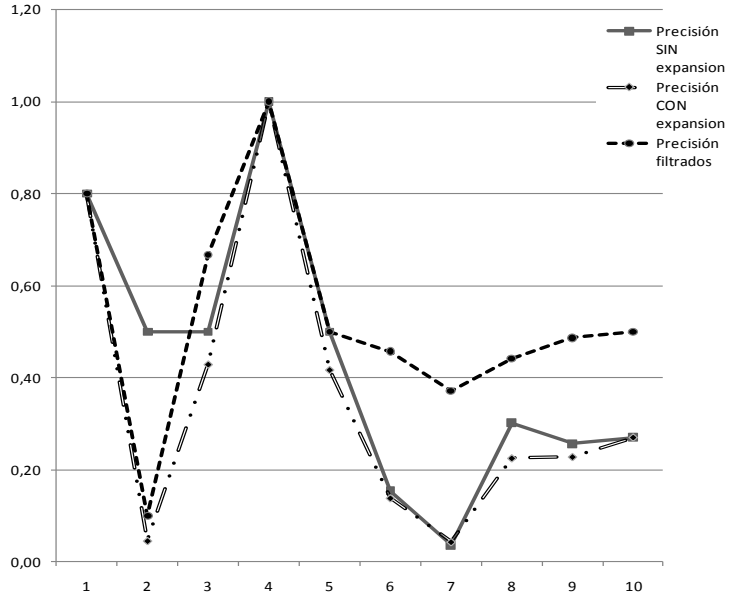

Figura 6: Precisión de las 12 consultas.

Se observa que en general los valores de Precisión Con Expansión están por debajo de los de

Precisión Sin Expansión**.** Esto se debe a que, como se comenta en el caso de estudio, el agregado de sinónimos e hiperónimos produce la recuperación de documentos que no son relevantes al interés del usuario. Sin embargo, la realización del filtrado y ordenado como posprocesamiento produce que los valores de precisión mejoren, y resulte una gráfica por arriba de las anteriores. Esto ocurre salvo en consultas como la 2 donde la cantidad de elementos recuperados con expansión es mucho mayor a la cantidad de documentos recuperados sin expansión.

Como resultado general, considerando los valores promedio de precisión éstos disminuyen en las consultas realizadas con expansión (de un valor promedio de 0,43 de la consulta sin expansión, se pasa a 0,36). Sin embargo, el posterior filtrado y ordenado de los resultados de la consulta expandida produce un interesante incremento de la precisión promedio a 0,53. Esto es, se logra un 23% de incremento en la precisión respecto a la consulta sin expansión.

6. Discusiones

En este trabajo se ha presentado la propuesta de un sistema de búsqueda semántica de objetos de aprendizaje en repositorios educacionales. Como recurso lingüístico se ha utilizado WordNet, una base de datos léxica en ingles, y como repositorio se ha elegido Ariadne, que cuenta con un número de objetos en su mayoría en ingles y bajo el estándar lom. La implementación se ha desarrollado bajo el framework 3.5 de .Net y se han utilizado los web services que provee el repositorio para hacer la conexión con el mismo.

De todos los repositorios analizados, Ariadne fue el más conveniente para la realización de las consultas. Tiene el mayor número de objetos (la mayoría en ingles), y su uso es gratuito.

El proceso de expansión de la consulta genero resultados más precisos dándole también al usuario, mediante el ranking, una idea de la relevancia de los objetos. Los hiperónimos presentaron ventajas en ciertos casos. Algunas veces restringieron de forma moderada los resultados obtenidos incrementando la precisión. Sin embargo otras veces condicionaron las consultas a tal punto que los resultados fueron nulos. Se observo que la precisión de los resultados aumentaba en mayor medida en aquellos casos donde la consulta recuperaba más de 20 elementos. En los casos donde el número de resultados era mínimo en general la precisión era la misma tanto para las consultas simples como para las expandidas.

Ciertas limitaciones pueden surgir por parte de los recursos lingüísticos utilizados. Por ejemplo, WordNet presenta falencias al momento de vincular términos altamente relacionados. Al observar la estructura de WordNet y las relaciones entre los elementos que la componen se verifican los vacíos existentes entre esos términos (por ejemplo *cancer* y *doctor* son palabras relacionadas en el ámbito de la medicina sin embargo es difícil encontrar una relación dentro de la estructura de WordNet). Asimismo aunque WordNet contiene una gran variedad de palabras comunes, no cubre el vocabulario específico de cada dominio. Además el reconocimiento de las frases ingresadas por el usuario está limitado a aquellas contenidas en WordNet.

Referencias

[1] Miller George A. (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

[2] Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press. Princeton University "About WordNet." WordNet. Princeton University. 2010. http://wordnet.princeton.edu

[3] Loupy, C., & El-Bèze, M. (2002). Managing synonymy and polysemy in a document retrieval system using WordNet, Proceedings of the LREC2002: Workshop on Linguistic Knowledge Acquisition and Representation.

[4] Philip Resnik. Disambiguating Noun Groups With Respect To WordNet Sense. In Proceedings Third Workshop on Very Large Corpora. 1995. pp. 54-68.

[5] Martínez Fernández, P.; García Serrano, A. (2002). Utilizando recursos lingüísticos para la mejora de la recuperación de información en la Web. Revista Iberoamericana de Inteligencia Artificial. VI/02, (16), 55-64.

[6] Ellen Voorhees. Using WordNet to disambiguate word senses for text retrieval. In Proceedings 16th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '93). ACM New York, NY, USA. 1993. pp 171-180.

[7] Dan I. Moldovan and Rada Mihalcea. Improving the search on the Internet by using WordNet and lexical operators. In IEEE Internet Computing. pp. 34-43. 1998.

[8] Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. Computational Linguistics 21(4), 543–565.

[9] Ariadne (Alliance of Remote Instructional Authoring and Distribution Networks for Europe), http://www.ariadne-eu.org