Distinción semántica de compuestos léxicos en Recuperación de Información

Anselmo Peñas, Julio Gonzalo y Felisa Verdejo

Dpto. Lenguajes y Sistemas Informáticos
ETS Ingenieros Industriales, UNED

{anselmo,julio,felisa}@lsi.uned.es

Resumen: La consideración de sintagmas no parece producir mejoras significativas en los modelos clásicos de Recuperación de Información. En general, se acepta que los criterios de proximidad proporcionan mejores resultados que un criterio de adyacencia. El trabajo que se presenta explora la hipótesis de que no todos los compuestos léxicos deben considerarse de la misma forma. Se propone un procedimiento automático de clasificación semántica de los compuestos léxicos de WordNet sobre la base de sus componentes, y se estudia cómo afecta esta distinción a la Recuperación de Información.
Palabras clave: Recuperación de información, compuestos léxicos.

1 Introducción

La consideración de sintagmas en los modelos de Recuperación de Información (IR) se ha estudiado profusamente en las últimas décadas. Por una parte, los sintagmas son relativamente fáciles de obtener y, por otra, no hay una única definición de sintagma en IR, considerándose como tales desde n-gramas validados estadísticamente hasta estructuras gramaticales obtenidas mediante técnicas lingüísticas. A este respecto, la conclusión general en la literatura es que el empleo de sintagmas identificados con técnicas lingüísticas no tienen por qué proporcionar mejores resultados en la recuperación que el uso de sintagmas identificados estadísticamente (Sparck Jones 1999).

Independientemente de cómo se hayan obtenido los sintagmas, los experimentos que han intentado corroborar si el uso de sintagmas permite o no mejorar la efectividad de la recuperación no han obtenido resultados concluyentes.

(Fagan 1989) evaluó la utilización de sintagmas identificados estadísticamente atendiendo a parámetros como el grado de proximidad o la frecuencia de aparición de las palabras componentes. La noción de proximidad relaja considerablemente el concepto de sintagma produciendo una mejora en la recuperación. Sus conclusiones fueron que no merece la pena considerar compuestos con más de dos constituyentes y que los compuestos no deben sustituir a sus componentes sino que deben considerarse ambos, tanto los compuestos como los elementos de los compuestos. Sin embargo, según Fagan, una de las razones para que los términos compuestos no puedan contribuir a mejorar la recuperación es simplemente que no aparecen en las consultas y que, por tanto, su presencia en los documentos se ignora.

(Croft 1991) también llega a la conclusión de que se debe pesar tanto el compuesto como sus elementos constituyentes.

(Krovetz 1997) volvió a estudiar el papel que puede desempeñar la indexación de sintagmas en la recuperación. Al igual que en experimentos anteriores, sus resultados muestran que la detección de colocaciones puede ser ligeramente beneficiosa para la recuperación, si bien resulta indispensable considerar las componentes de la colocación asignándoles también un peso parcial.

También en las experiencias TREC se ha abordado la posibilidad de considerar sintagmas como parte de una aplicación exhaustiva de las técnicas NLP en IR (Strzalkowski 1997), (Strzalkowski 1998), (Strzalkowski 1999). Sin embargo, estas experiencias tampoco han conseguido mejorar la recuperación.

Del trabajo de (Pickens 2000), sin embargo, se puede concluir que la adyacencia de las palabras de la consulta en un documento supone una evidencia positiva en favor de la relevancia del documento y, por tanto, que una

consideración adecuada de sintagmas debería llevar a un aumento en la precisión de la recuperación.

Ninguno de los trabajos mencionados distingue tipos diferentes de sintagmas sino que, implícitamente, asumen que el comportamiento de todos ellos es similar en IR. En este trabajo se estudia la hipótesis de que no todos los sintagmas presentan el mismo comportamiento en IR, centrando el estudio en los compuestos léxicos de WordNet. En el apartado 2 se expone una clasificación de compuestos léxicos bien conocida en inglés, atendiendo a criterios semánticos. El apartado 3 propone un método de clasificación automática de los compuestos de WordNet, atendiendo a los criterios expuestos. El apartado 4 muestra una serie de experimentos de recuperación dirigidos a evaluar cómo afecta la distinción semántica de compuestos a la Recuperación de Información.

2 Tipos de compuestos léxicos

Antendiendo a la presencia de relaciones de hiponimia/ hiperonimia entre el compuesto y sus componentes, cabe realizar la siguiente distinción semántica de compuestos nominales en inglés:

2.1 Compuesto Endocéntrico

Semánticamente, se distingue un compuesto endocéntrico porque es un hipónimo de una de sus componentes (generalmente la última), desempeñando el resto de componentes el papel de modificadores. Por ejemplo, *armchair* es un tipo de *chair* mientras que *arm* juega el papel de modificador de *chair*; *"toothed whale"* es un tipo de *whale*, es decir, "*toothed whale"* es un hipónimo de *whale.* Este tipo de compuestos heredan las propiedades sintácticas (género, número, etc.) de una de sus componentes (en inglés, generalmente la última).

2.2 Compuesto Exocéntrico

Un compuesto exocéntrico se distingue semánticamente porque es un hipónimo de un elemento no expresado en el compuesto. Por ejemplo, *hatchback* es un tipo de *car*, o "*mentally retarded*" es un tipo de *people.* Sin embargo, esta diferencia semántica respecto a los compuestos endocéntricos no se presenta a nivel gramatical. Al igual que los compuestos endocéntricos, los compuestos exocéntricos heredan las propiedades sintácticas de una de sus componentes (generalmente la última).

2.3 Compuesto Aposicional

Semánticamente, un compuesto aposicional es un hipónimo de varias de sus componentes de forma simultánea. Por ejemplo, *"folk song"* es un tipo tanto de *music* como de *folk*. En este caso no se pierde mucha información si estos compuestos se tratan como compuestos endocéntricos. Por ejemplo, en el caso de *girlfriend* no hay mucha diferencia entre una representación semántica de *"both, girl and friend"* o sólo *"a friend who is a girl"*. En este tipo de compuestos se considera que, sintácticamente, una de las componentes es la principal.

2.4 Compuesto Copulativo (o Dvandva)

Los compuestos copulativos son el resultado de combinar entidades separadas para referenciar a una única entidad como, por ejemplo, *pantyhose* (*pantis, medias*). Salvo el ejemplo, se trata de nombres propios. En este tipo de compuestos no siempre se distingue si hay una componente principal.

Las entidades representadas por los compuestos copulativos no están incluídas en las jerarquías de WordNet y, por ello, no serán tratadas en este trabajo. En realidad se trata de un problema de Reconocimiento de Entidades en el que no entraremos aquí.

3 Propuesta de clasificación automática de compuestos léxicos de WordNet

De los cuatro tipos de compuestos definidos en el apartado anterior, sólo los compuestos copulativos no están en WordNet. El resto sí están presentes en WordNet permitiendo estudiar las relaciones de hiperonimia/ hiponimia y sinonimia entre el compuesto y sus componentes. Sobre la base de estas relaciones vamos a proponer un método automático que permite clasificar los compuestos léxicos contenidos en WordNet.

Trabajando con WordNet, el proceso de clasificación tiene que tener en cuenta que una componente de un compuesto léxico puede ser a su vez otro compuesto léxico. Por ejemplo, *Atlantic bottlenose dolphin* es un hipónimo en WordNet de *bottlenose dolphin*. También debe tenerse en cuenta que, en WordNet, hay ocasiones en las que una componente es sinónimo del propio compuesto. Por ejemplo, *primary* es sinónimo de *primary election.* Por

esta razón, la clasificación de compuestos no debe considerar únicamente hiperónimos, sino también sinónimos.

En WordNet 1.5 hay más de 56.000 compuestos léxicos. De ellos, tan sólo 2860 son polisémicos (5%). Estos compuestos polisémicos han sido excluidos del proceso de clasificación que se describe en los siguientes subapartados.

3.1 Clasificación de compuestos endocéntricos

La forma de distinguir estos compuestos es comprobar que una de sus componentes es hiperónimo del compuesto. Puede ocurrir que se trate del hiperónimo inmediatamente superior, que se trate de otro más alto en la jerarquía, incluso que se den ambos casos simultáneamente como ilustra el ejemplo (*Atlantic bottlenose dolphin*):

```
1 sense of atlantic bottlenose dolphin

Sense 1
Atlantic bottlenose dolphin, Tursiops truncatus
      => bottlenose dolphin, bottle-nosed dolphin, bottlenose
         => dolphin
            => toothed whale
               => whale
                  => cetacean, cetacean mammal, blower
                     => aquatic mammal
```

Los ejemplos muestran la secuencia de hiperónimos del compuesto hasta el nivel más alto de la jerarquía. Así, *Atlantic bottlenose dolphin* tiene como hiperónimo inmediato a *bottlenose dolphin,* a *dolphin* como hiperónimo del siguiente nivel, etc. Para clasificar un compuesto como *compuesto endocéntrico*, los hiperónimos que coincidan con alguna de las componentes deben pertenecer a la misma rama de la jerarquía. En otro caso se trataría de compuestos aposicionales como se verá más adelante. Otros ejemplos de compuestos endocéntricos son los siguientes:

```
1 sense of primary election

Sense 1
primary, primary election
      => election
         => vote
            => group action
               => act, human action, human activity
```

En el caso de *primary election,* únicamente aparece la componente *election* como hiperónimo, lo que determina que el compuesto sea *endocéntrico*. Si hubiera aparecido también *primary* como hiperónimo pero por otra rama diferente, entonces *primary election* se habría clasificado como *compuesto aposicional*.

El siguiente ejemplo muestra un caso muy similar: *jury* es hiperónimo de *grand jury* y además es el único, por lo que el compuesto se clasifica como *endocéntrico*.

```
1 sense of grand jury

Sense 1
grand jury
      => jury
         => body
            => gathering, assemblage
               => social group
                  => group, grouping
```

El siguiente ejemplo muestra un caso en el que la componente hiperónimo no está en el nivel inmediatamente superior. Así, para identificar *department* como componente hiperónima de *purchasing department* hay que ascender dos niveles en la jerarquía.

```
1 sense of purchasing department

Sense 1
purchasing department
      => business department
         => department
            => division
               => administrative unit
                  => unit
                     => organization
                        => social group
                           => group, grouping
```

El caso de *alloy steel* muestra un caso en el que ambas componentes son hiperónimas del compuesto. Sin embargo, y aunque el compuesto tiene dos ramas de hiperónimos, ambas componentes están en la misma rama. Esto determina que el compuesto se clasifique como endocéntrico y no como aposicional.

```
1 sense of alloy steel

Sense 1
alloy steel
      => steel
         => alloy
            => mixture
               => substance, matter
                  => object, inanimate object, physical object
                     => entity
            => metallic element, metal
               => chemical element, element
                  => substance, matter
                     => object, inanimate object, physical object
                        => entity
```

En algunos casos, los compuestos no tienen ninguna componente hiperónima pero, sin embargo, existe alguna componente que es sinónimo del compuesto. Estos casos también se van a clasificar como compuestos endocéntricos. Este es el caso de *bank account* cuyas componentes no son hiperónimos del compuesto pero que sin embargo tiene a *account* como sinónimo.

```
1 sense of bank account

Sense 1
account, bank account
      => fund, monetary fund
         => money
            => medium of exchange, monetary system
               => asset
                  => possession
```

También puede darse el caso en que una componente es sinónimo e hiperónimo del compuesto simultáneamente. Por ejemplo, el compuesto *electric drill* tiene a su componente *drill* tanto como sinónimo como hiperónimo.

```
1 sense of electric drill

Sense 1
drill, electric drill
      => power drill
         => power tool
            => machine
               => device
                  => instrumentality, instrumentation
                     => artifact, artefact
                        => object, inanimate object, physical object
                           => entity
      => drill
         => tool
            => implement
               => instrumentality, instrumentation
                  => artifact, artefact
                     => object, inanimate object, physical object
                        => entity
```

Lo mismo ocurre con *kentuky yellowwood* que tiene a su componente *yellowwood* como sinónimo y como hiperónimo.

```
1 sense of kentucky yellowwood

Sense 1
Kentucky yellowwood, gopherwood, Cladrastis lutea, Cladrastis
kentukea, yellowwood
      => angiospermous yellowwood
         => yellowwood, yellowwood tree
            => tree
               => woody plant, ligneous plant
                  => vascular plant, tracheophyte
                     => plant, flora, plant life
                        => life form, organism, being, living thing
                           => entity
```

3.2 Clasificación de compuestos aposicionales

La forma de detectar un compuesto aposicional es comprobar que más de una componente diferente es hiperónimo del compuesto pero en jerarquías diferentes (herencia múltiple). Por ejemplo, *aspirin powder* tiene a su componente *aspirin* como hiperónimo en una rama y a su componente *powder* en otra.

```
1 sense of aspirin powder

Sense 1
aspirin powder, headache powder
      => aspirin, acetylsalicylic acid, Bayer, Empirin
         => analgesic, anodyne, painkiller, pain pill
            => medicine, medication, medicament, medicinal drug
               => drug
                  => artifact, artefact
                     => object, inanimate object, physical object
                        => entity
      => powder
         => toiletry, toilet article, toiletries
            => instrumentality, instrumentation
               => artifact, artefact
                  => object, inanimate object, physical object
                     => entity
      => medicine, medication, medicament, medicinal drug
         => drug
            => artifact, artefact
               => object, inanimate object, physical object
                  => entity
```

Lo mismo ocurre con el compuesto aposicional *folk song* en el que tanto *folk* como *song* son hiperónimos en jerarquías diferentes.

```
1 sense of folk song

Sense 1
folk song, folk ballad
      => folk music, ethnic music, folk
         => music
            => art, fine art
               => creation
                  => artifact, artefact
                     => object, inanimate object, physical object
                        => entity
      => song
         => musical composition, opus, composition, piece, piece
of music
            => music
               => art, fine art
                  => creation
                     => artifact, artefact
                        => object, inanimate object, physical object
                           => entity
```

En los compuestos aposicionales se van a incluir aquellos en los que una componente es sinónimo del compuesto y otra componente distinta es un hiperónimo. Por ejemplo, *1st-class mail* tiene a *1st-class* como sinónimo y a *mail* como hiperónimo. Lo mismo ocurre con *abductor muscle* o *abrasive material*.

```
1 sense of 1st-class mail

Sense 1
first-class, 1st-class, first-class mail, 1st-class mail, priority mail
      => mail
         => message
            => communication
               => social relation
                  => relation
                     => abstraction
```

```
1 sense of abductor muscle

Sense 1
abductor, abductor muscle
      => skeletal muscle
         => muscle, musculus
            => contractile organ
               => organ
                  => body part
                     => part, piece
                        => entity
```

```
1 sense of abrasive material

Sense 1
abrasive, abradant, abrasive material
      => material, stuff
         => substance, matter
            => object, inanimate object, physical object
               => entity
```

3.3 Clasificación de compuestos exocéntricos

Para clasificar un compuesto como *exocéntrico* hay que comprobar que ninguna de sus componentes es sinónimo o hiperónimo del mismo. Esto ocurre, por ejemplo con *fisher cat*, *man and wife* o *mentally retarded*, en los cuales ninguna de sus componentes es un hiperónimo del compuesto.

```
1 sense of fisher cat

Sense 1
fisher, pekan, fisher cat, black cat, Martes pennanti
      => marten, marten cat
         => musteline mammal, mustelid, musteline
            => carnivore
               => placental mammal, eutherian, eutherian mammal
                  => mammal
                     => vertebrate, craniate
                        => chordate
                           => animal, animate being, beast, brute, creature, fauna
                              => life form, organism, being, living thing
                                 => entity
```

```
1 sense of man and wife

Sense 1
marriage, married couple, man and wife
      => family, family unit
         => kin, kin group, kinship group, kindred, clan, tribe
            => social group
               => group, grouping
```

```
1 sense of mentally retarded

Sense 1
mentally retarded
      => people
         => group, grouping
```

4 Clasificación automática de compuestos léxicos aplicada a la Recuperación de Información

En esta propuesta se van a distinguir dos grupos de compuestos de cara a la Recuperación de Información:

- Compuestos endocéntricos y aposicionales.
- Compuestos exocéntricos.

A partir de esta distinción se mostrarán los experimentos dirigidos a evaluar cómo afecta esta distinción a la Recuperación de Información.

Los lenguajes de consulta de la mayoría de motores de búsqueda incluyen operadores de proximidad y de adyacencia. De esta manera, no es necesario detectar ni extraer los compuestos léxicos de los textos, basta con imponer restricciones sobre las palabras de los compuestos léxicos de la consulta, de acuerdo con la tipología propuesta en este trabajo.

4.1 Distinción de compuestos endocéntricos y aposicionales en IR

Los compuestos aposicionales pueden considerarse un caso particular de los compuestos endocéntricos en los que son varias y no una las componentes hiperónimas o sinónimas del compuesto.

En ambos casos, las componentes no pierden su significado, sino que modifican un sentido nuclear y, por tanto, es preferible no imponer una restricción de adyacencia sino únicamente de proximidad o, incluso, mantenerlas separadas. No hay criterios claros ni determinantes para decidir si la componente nuclear debe pesarse más o menos que las demás. Si queremos centrar la búsqueda en el *tema* general de la consulta, parece conveniente pesar más la componente nuclear del compuesto. Esta componente se identificará en el propio proceso de clasificación puesto que se trata de la componente hiperónima del compuesto.

4.2 Distinción de compuestos exocéntricos en IR

En el caso de los compuestos exocéntricos las componentes pierden su significado para crear uno nuevo resultante. Por esta razón, las componentes de un compuesto exocéntrico no deben considerarse por separado, sino que parece conveniente imponer la restricción de adyacencia sobre las palabras del compuesto. WordNet 1.5 tiene 19.284 compuestos exocéntricos sumando todas las categorías (nombres, adjetivos, verbos y adverbios). Esto supone que el 34% de los compuestos léxicos de WordNet pueden clasificarse como *exocéntricos*. Sin embargo, este dato no quiere decir que sean muy frecuentes en los textos, ni mucho menos en las consultas.

4.3 Definición del experimento

El experimento tiene como objetivo comparar precisión y cobertura de la recuperación cuando se distinguen compuestos léxicos.

La colección de prueba utilizada es OHSUMED que tiene 380Mb de documentos y 101 consultas en el dominio médico. Debido a que la clasificación de compuestos expuesta en el apartado anterior se realiza a partir de WordNet, su utilidad en Recuperación de Información depende de lo bien que WordNet cubra el dominio de búsqueda. En este caso, la colección OHSUMED resulta apropiada para el experimento porque las subjerarquías de WordNet relativas al dominio médico son bastante ricas y, por tanto, se espera que la recuperación se vea afectada por la distinción de compuestos.

El motor de búsqueda empleado ha sido INQUERY (Callan 1992). Las colecciones se han indexado en formato texto original sin *stemming.* Las consultas se han procesado de diferente manera para cada uno de los experimentos, de acuerdo con el tratamiento descrito anteriormente. Los experimentos que se han comparado son los siguientes:

1. *Sin compuestos*. Las consultas no se han procesado en ningún sentido salvo para adecuarlas al lenguaje de consulta del motor de búsqueda.
2. *Adyacencia*. A todos los compuestos detectados en las consultas se les ha impuesto la restricción de adyacencia, es decir, las palabras del compuesto deben encontrarse en el texto exactamente en la misma secuencia, y sin posibilidad de considerar las componentes aisladas. El lenguaje de consulta de INQUERY permite realizar este tratamiento mediante el operador *#ws, window size*, obligando a que el tamaño de la ventana en que deben aparecer las palabras sea igual al número de palabras del compuesto.
3. *Proximidad*. En este caso, en lugar de exigir la adyacencia de las palabras del compuesto, se pide que aparezcan en un entorno próximo, pero además otorgando un crédito parcial a la ocurrencia aislada de las componentes en el texto. El lenguaje de consulta de INQUERY permite realizar esta operación mediante el operador *#phrase* aplicado al compuesto.
4. *Adyacencia en exocéntricos y proximidad en el resto de compuestos*. En este caso, a los compuestos exocéntricos se les impone la restricción de adyacencia (operador *#ws* con tamaño igual al número de componentes), mientras que al resto de compuestos en las consultas se les aplica el operador de proximidad (*#phrase*).
5. *Restricción de adyacencia sólo para compuestos exocéntricos*.
6. *Restricción de adyacencia con sobrepeso sólo para compuestos exocéntricos*. En este caso, se distinguen únicamente compuestos exocéntricos pero, además, se les aplica un sobrepeso en la consulta. El lenguaje de consulta de INQUERY permite realizar esta operación gracias al operador #+.

4.4 Realización del experimento y resultados

La *Tabla 1* muestra la precisión obtenida en 10 puntos de *recall* para cada uno de los experimentos:

1. *Sin compuestos*. La precisión media en los 10 puntos de recall es del 19.2%.
2. *Adyacencia*. En este experimento, la precisión media en la recuperación baja a 15.8%, lo que supone una pérdida del 17.7%. En este caso no se han considerado de forma aislada las componentes de los compuestos y esto ha provocado una pérdida de precisión.
3. *Proximidad*. La precisión media obtenida al utilizar el operador de proximidad sube al 18.4% pero no llega a la precisión obtenida cuando no se consideran compuestos (19.2%). Esto indica que el crédito parcial asignado a las componentes no es

	Precisión (101 consultas)					
Recall	Sin considerar compuestos	Adyacencia	Proximidad	Adyacencia exocentricos, Proximidad resto	Adyacencia exocéntricos	Adyacencia exocéntricos con sobrepeso
10	44.4	40.9	43.3	43.1	44.5	44.5
20	35.7	32.3	35.6	35.6	37.3	37.4
30	29.0	23.4	27.5	27.3	29.3	29.4
40	23.4	19.0	22.1	22.0	23.2	23.3
50	19.7	15.1	18.6	18.6	19.9	19.9
60	13.8	11.1	12.7	12.7	13.6	13.7
70	10.4	7.3	9.5	9.5	10.2	10.2
80	7.7	5.1	7.0	7.1	7.4	7.4
90	4.9	2.7	4.4	4.3	4.7	4.7
100	3.0	1.5	2.9	2.9	3.0	3.0
Media	19.2	15.8	18.4	18.3	19.3	19.4

Tabla 1. Distinción de compuestos en Recuperación de Información

suficientemente alto con el operador #phrase.

4. *Adyacencia en exocéntricos y proximidad en el resto de compuestos*. En este experimento se distinguen compuestos exocéntricos obligando a la adyacencia de sus componentes. Sin embargo, al resto de compuestos se les sigue imponiendo una restricción de proximidad de las componentes que impide un aumento de precisión. La precisión media prácticamente coincide con la anterior, siendo del 18.3%.
5. *Restricción de adyacencia sólo para compuestos exocéntricos*. En este caso sólo se consideran compuestos exocéntricos imponiendo la restricción de adyacencia sobre sus componentes. El efecto es que la precisión media vuelve a los niveles de una recuperación sin considerar compuestos, siendo del 19.3%, apenas una décima por encima.
6. *Restricción de adyacencia con sobrepeso sólo para compuestos exocéntricos*. En este caso, similar al anterior, se le otorga más peso a los compuestos exocéntricos que a cualquier otro término de la consulta. El efecto es que la precisión media sube al 19.4%, tan sólo dos décimas más que una recuperación sin considerar compuestos.

5 Conclusiones

En este trabajo se ha propuesto una forma de clasificar los compuestos léxicos de WordNet atendiendo a criterios semánticos, y se ha mostrado un estudio preliminar de cómo afecta esta distinción a la recuperación de información.

Los resultados confirman la hipótesis de que los compuestos exocéntricos no se comportan igual que los endocéntricos y aposicionales. Mientras que la consideración de compuestos endocéntricos produce una pérdida de precisión, la consideración únicamente de compuestos exocéntricos produce una mejora aunque muy poco significativa.

Este comportamiento de los compuestos exocéntricos es lógico puesto que las componentes no mantienen un significado parcial y, por tanto, la consideración de las componentes por separado conduce a resultados incorrectos.

Sin embargo, la diferencia entre los valores de precisión obtenidos es demasiado reducida como para emitir un resultado concluyente. Esto se debe fundamentalmente a que el número de compuestos léxicos en las consultas de la colección OHSUMED resulta muy reducido. Sólo el 13% de las consultas contiene algún compuesto de WordNet, y sólo el 7% de las consultas contienen un compuesto exocéntrico.

6 Trabajo futuro

Aunque debido a la escasez de compuestos en las consultas de OHSUMED las diferencias en los resultados hayan sido mínimas, el hecho de que un sobrepeso sobre los compuestos exocéntricos eleve algo la precisión media de la recuperación parece indicar que resulta conveniente su consideración, sugiriendo la conveniencia de una investigación más profunda. Precisamente, uno de los trabajos futuros debe dirigirse a determinar el valor

óptimo de este sobrepeso.

Si bien resulta interesante repetir el experimento considerando únicamente las consultas que contienen compuestos exocéntricos, es necesario realizar los experimentos con consultas más extensas y con mayor posibilidad de contener compuestos exocéntricos. En este sentido, las colecciones TREC pueden resultar apropiadas utilizando como consulta la sección *narrative* de los *topics*. Otra posibilidad podría ser estudiar la detección de compuestos exocéntricos en un marco de *pseudo relevance feedback.*

Por otra parte, en los experimentos mostrados aquí, únicamente se han detectado los compuestos en las consultas, no en los textos. Sin embargo, resulta de interés estudiar como afectaría a la recuperación la indexación de los compuestos exocéntricos en los textos. De esta manera, por ejemplo, una consulta con *fisher* no recuperaría un texto con *fisher_cat.*

Respecto a la precisión del proceso automático de clasificación de compuestos, es necesario estudiar la lista de compuestos exocéntricos. Como WordNet es una red semántica construída manualmente, cabe confiar en sus relaciones semánticas. Esto implica que si un compuesto tiene una componente hiperónima entonces el compuesto es endocéntrico o aposicional con suficiente seguridad. Sin embargo, la clasificación de compuestos que no tienen componentes hiperónimas (candidatos a ser exocéntricos) no es tan fiable, ya que la falta de hiperónimos puede deberse a una falta de conceptos en la red y no a que realmente se trate de un compuesto exocéntrico.

```
1 sense of abstract artist

Sense 1
abstractionist, abstract artist
      => painter
         => artist, creative person
            => creator
               => person, individual, someone, mortal, human, soul
                  => life form, organism, being, living thing
                     => entity
                  => causal agent, cause, causal agency
                     => entity
```

Por último, resulta interesante estudiar las posibilidades que introduce situar un compuesto en la red semántica de WordNet. Por ejemplo, en el caso de que una componente hiperónima no sea del nivel inmediatamente superior de la jerarquía de WordNet, puede ser interesante añadir a la consulta, a modo de expansión, los términos de los synsets de niveles intermedios. Por ejemplo, en el caso de *"abstract artist*", *artist* es hiperónimo de segundo nivel y podría añadirse *painter*:

7 Agradecimientos

Este trabajo ha sido parcialmente financiado por el Ministerio de Ciencia y Tecnología a través del proyecto Hermes (TIC2000-0335-C03-01).

8 Referencias

Callan, J. Croft B. and Harding S. The INQUERY retrieval system. Proceedings of the 3rd International Conference on Database and Expert Systems applications; 1992.

Croft, W. B. Turtle H. R. and Lewis D. D. The use of phrases and structured queries in information retrieval. Proceedings of 14th SIGIR Conference on Research and Development in Information Retrieval. 1991; 32-45.

Fagan, J. L. The effectiveness of a non-syntactic approach to automatic phrase indexing for document retrieval. Journal of the American Society for Information Science. 1989; 40(2):115-132.

Krovetz, R. Homonymy and polysemy in Information Retrieval. ACL/EACL'97; 1997.

Pickens, J. and Croft W. B. An Exploratory analysis of Phrases in Text Retrieval. Proceedings of RIAO 2000 Conference, Paris. 2000; 1179-1195.

Sparck Jones, K. What is the Role of NLP in Text Retrieval? Natural Language Information Retrieval, Ed. T. Strzalkowski, Kluwer Academic Publishers. 1999.

Strzalkowski, T. Natural language Processing Information Retrieval. Kluwer, Boston, MA. 1999.

Strzalkowski, T. Lin F. Pérez-Carballo J. and Wang J. Natural Language Information Retrieval: TREC-6 Report. Proceedings of TREC-6 Conference. 1997.

Strzalkowski, T. Stein G. Wise G. B. Pérez-Carballo J. Tapanainen P. jarvinen T. Voutilainen A. and Karlgren J. Natural Language Information Retrieval: TREC-7 Report. Proceedings of TREC-7 Conference. 1998.