Internet como fuente de información léxica: extracción de etiquetas de dominio y detección de nuevos sentidos

Celina Santamaría Julio Gonzalo Felisa Verdejo
Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia (UNED)
{celina,julio,felisa}@lsi.uned.es

Resumen Describimos un algoritmo que combina información léxica (extraída de WordNet 1.6) con información en Internet (directorios de Altavista) para caracterizar automáticamente los sentidos de una palabra con etiquetas de dominio y, al mismo tiempo, detectar y describir nuevos sentidos relevantes en Internet.

Esta información puede utilizarse, entre otras cosas, para enriquecer bases de datos léxicas, para facilitar la extracción de corpora anotados semánticamente y derivados de Internet (como material de entrenamiento para sistemas de anotación semántica), o para agrupar sentidos (de dominio coincidente) cuando las distinciones semánticas son demasiado finas para las aplicaciones en que se usan.

1 Introducción

Internet, como fuente casi ilimitada de información textual, es un recurso prometedor para muchas áreas de Ingeniería Lingüística y, en particular, para el enriquecimiento de bases de datos léxicas. En este artículo presentamos un algoritmo mediante el cual extraemos, a partir de la información de directorios de Altavista[3], etiquetas de dominio que caracterizan los sentidos de los nombres presentes en WordNet[8]. Al mismo tiempo, el algoritmo detecta nuevos sentidos predominantes en Internet, que pueden usarse para añadir nuevas entradas a WordNet o para filtrar y refinar el material obtenido en la red. Algunas de sus posibles aplicaciones son:

- **Adquisición automática de corpora etiquetados semánticamente.** Un precedente de este tipo de técnicas es [6]. En este trabajo se utiliza información extraída de WordNet acerca de cada sentido w_i, para generar consultas a la red, de forma que los documentos obtenidos contengan ocurrencias de w con una elevada probabilidad de pertenecer al sentido w_i. Estas ocurrencias y su contexto, a priori, pueden servir como material para entrenar sistemas supervisados de resolución de la ambigüedad léxica (Word Sense Disambiguation, WSD). Sin embargo, pese a que los autores atribuyen una alta precisión a este método, los datos extraídos no resultan útiles como material de entrenamiento en la comparativa realizada en [2].

  Una causa razonable se encuentra en que, dependiendo de la palabra en cuestión, es muy posible que existan sentidos predominantes en Internet, pero ausentes de WordNet, que interfieran negativamente en la obtención de ocurrencias con los sentidos deseados. Por ejemplo, es difícil obtener resultados apropiados para los sentidos de *oasis* sin filtrar a priori aquellos documentos que tratan sobre el grupo musical Oasis; o sobre *tiger* sin descartar documentos sobre el golfista *Tiger Woods*.

  Nuestro algoritmo detecta este tipo de interferencias, permitiendo mejorar la precisión sobre palabras con este patrón de comportamiento.

  Ora característica destacable es que los ejemplos están asignados a un dominio (directorio de Altavista), lo que permite clasificar los datos de entrenamiento y comprobar si son compatibles con el dominio sobre el que se está probando el sistema.

  La información de dominio extraída de la red puede en general beneficiar a otros métodos en los que se utiliza la web para obtener información semántica, como el de [1], en el cual la WWW se emplea para enriquecer una ontología con etiquetas temáticas.

- **Agrupación de sentidos.**
  Los diccionarios electrónicos y las bases de datos léxicas como WordNet o LDOCE tienden a presentar una granularidad excesivamente fina para la mayoría de las aplicaciones en NLP; parece razonable, para muchas utilidades, identificar aquellos sentidos que pertenecen al mismo dominio [5]. Si se pueden caracterizar dos sentidos con la misma categoría de Altavista, probablemente no necesitan ser distinguidos para ciertas aplicaciones. Por ejemplo, en tareas de búsqueda de información, distinguir sentidos de dominios similares no mejora la precisión, mientras que distinguir sentidos de dominios distintos si puede hacerlo.
- **Aplicaciones de traducción.**
  Tanto en Traducción Automática como en Búsqueda de Información Multilingüe, la detección de sentidos ausentes en el diccionario bilingüe es necesaria para evitar traducciones incorrectas. En el caso de búsqueda de información multilingüe, un error de traducción puede tener un impacto notable en la precisión cuando se trata de consultas de tamaño reducido. Por ejemplo, sería un error traducir *tiger* por *tigre* si la consulta versa sobre el golfista.
  En general, cualquier aplicación de Ingeniería Lingüística se beneficiará de una cobertura léxica mayor y una adaptación automática al dominio, sea éste Internet (por ejemplo, en traductores como www.babelfish.com) o un dominio más reducido. Otro ejemplo puede ser el de los enfoques de búsqueda de información basados en indexación semántica como [7, 4], en los que es crucial tanto disponer de una cobertura adecuada como evitar excesivas distinciones de sentido si pertenecen a un mismo dominio.

En la sección 2 describimos el algoritmo propuesto. En la sección 3 evaluamos sus resultados para un subconjunto reducido de nombres presentes en WordNet. Finalmente, en la sección 4 extraemos algunas conclusiones sobre esta aproximación.

2 Algoritmo

En resumen, nuestro algoritmo toma un nombre (de WordNet 1.6) como entrada. Genera entonces un conjunto de consultas, una por cada sentido presente en WordNet. Lanza cada consulta sobre Altavista, y clasifica las directorios resultantes como a) apropiados como etiqueta de dominio para algún sentido existente en WordNet; b) apropiados para describir un nuevo sentido; c) ruido, que debe ser filtrado mediante refinamiento de las consultas. Este es el proceso detallado:

1. La entrada inicial para el algoritmo es un nombre w presente en WordNet 1.6. Para cada sentido w_i de esa palabra en WordNet se genera automáticamente una consulta q_i, en la que w se incluye como término obligatorio. El resto de la consulta está formado por los sinónimos del correspondiente synset de WordNet, junto con los hiperónimos directos (todos ellos incluidos como términos opcionales), y los sinónimos de los demás sentidos de w como términos negados (es decir, que no deben aparecer en los documentos que devuelva Altavista). Por ejemplo , para *tiger* se generan las siguiente consultas:

   ```
   q1= [+tiger person individual
   someone somebody mortal human
   soul -"Panthera tigris" ]
   ```

   (*tiger 1 – a fierce or audacious person; "he's a tiger on the tennis court"; it aroused the tiger in me"*)

   ```
   q2= [+tiger "Panthera tigris"
   "big cat" cat ]
   ```

   (*tiger2, Panthera tigris – (large feline of forests in most of Asia having a tawny coat with black stripes; endangered)*

2. Cada consulta se envía a los directorios de Altavista. Se trata de una vasta colección de documentos, organizada en una jerarquía de directorios que clasifican la información por dominios y subdominios. A partir de aquí, se considera una lista formada por todos los directorios que incluyen explícitamente la palabra inicial w. Para cada subdirectorio d, se forma una lista de palabras l_d lematizando todos los términos que aparecen en el directorio (exceptuando las palabras vacías). Nos referiremos a estas listas como *dominios*.

Por ejemplo, la lista de dominios obtenida para *tiger* se forma lematizando dominios como:

```
sports/all sports/golf/professional
 golfers/pga golfers/woods, tiger

library/sciences/animals & wildlife
 /mammals/wildcats/tigers

entertainment/movies/movies by
 genre/drama/epics/crouching tiger
```

3. Para cada sentido w_j, se forma una lista de palabras l_j empleando todos los nombres presentes en la cadena de hiperónimos, asi como los hipónimos directos, merónimos, holónimos y términos coordinados de w_j,obtenidos de WordNet. l_j se utiliza como descripción del sentido w_j.

4. Para cada descripción de sentido l_j, se establece una comparación entre sus términos y los términos de cada dominio l_d. Esta comparación se basa en la hipótesis de que los términos de un dominio adecuado para un sentido particular w_i estarán correlacionados con w_i mediante las relaciones semánticas de WordNet.

   Cuando una descripción de sentido l_j y una descripción de dominio l_d comparten al menos un término, d es asociado al sentido w_j. Si coinciden el sentido original de la consulta i (que retornó d) con el sentido j ($i = j$), llamamos a d "etiqueta positiva"para el sentido i. En caso contrario, d se considera "etiqueta negativa"para el sentido i. Si l_d no comparte ningún término con ningún sentido, llamamos "ruido" al dominio d. Las etiquetas clasificadas como ruido se descartarán, o bien se reconocerán como nuevos sentidos.

   A continuación se describe el tratamiento para cada una de estas situaciones:

5. **Tratamiento de etiquetas**. Se presentan dos posibilidades:

   (a) El algoritmo asocia una misma etiqueta a varios sentidos. En tal caso, estos sentidos se consideran candidatos para agrupamiento. La hipótesis subyacente es que no es preciso distinguir dos sentidos que aparecen asociados al mismo dominio, al menos en un subconjunto apreciable de aplicaciones.

   (b) Una etiqueta resulta asociada a un único sentido. En este caso, se utilizan primero dos filtros adicionales:

      i. detectamos aquellos casos en los que un término del descriptor asociado contiene al término anterior, como ocurre en esta etiqueta asignada a un sentido de *lion*:

      ```
      library/sciences/animals
      & wildlife/mammals/mammals
      a-z/mammals t-z/tamarins/
      golden lion tamarin
      ```

      golden lion tamarin es una especialización de *tamarin*, y por tanto *lion* está actuando como modificador, y no en relación con los sentidos de *lion* en WordNet. En estos casos, se descarta la etiqueta y se clasifica como ruido.

      ii. Otro filtro, más débil que el anterior, es descartar compuestos en los que el nombre preceda a otra palabra, como en:

      ```
      personal/kids/
      arts & entertainment/
      movies/animals/lion king
      ```

      En general se trata de adjetivos, y se descartan para favorecer la precisión (sacrificando, a priori, la cobertura). De nuevo, en estos casos la etiqueta se reclasifica como ruido.

   Si la etiqueta pasa ambos filtros, se generan dos tipos de información. En primer lugar, asociamos una etiqueta de dominio a ese sentido, enriqueciendo así WordNet con información contextual. Por ejemplo, el algoritmo genera:

```
tiger 2 →

[library, sciences, animals &
  wildlife, mammals, wildcats]
```

En segundo lugar, Generamos una familia de nuevas consultas, refinadas. En el caso de etiqueta positiva para el sentido i se añaden los

términos de la etiqueta l_d a la consulta original q_i. Para el resto de sentidos, los términos de la etiqueta l_d se añaden como términos negativos para la búsqueda. Por ejemplo:

```
[+tiger "Panthera tigris"
"big cat" cat library sciences
"animals & wildlife" mammals
wildcats]

[+tiger person individual
someone somebody mortal
human soul -"Panthera tigris"
-library -sciences
-"animals & wildlife" -mammals
-wildcats]
```

Los sentidos previamente colapsados con i no se consideran.

6. **Tratamiento del ruido**. Se consideran dos aplicaciones:

   (a) **Refinamiento de consultas**: Para todos los sentidos, los elementos de l_d se añaden como términos negativos con la intención de mejorar la precisión de los documentos asociados a cada sentido. Por ejemplo, los términos de la categoría Altavista correspondiente a *Tiger Woods* se añaden negativamente a las consultas para *tiger*:

```
tiger #1 ->
[+tiger person individual
someone somebody mortal
human soul -"Panthera tigris"
-sports -"all sports" -golf
-"professional golfers"
-"pga golfers" -"woods, tiger"]

tiger #2 ->
[+tiger "Panthera tigris"
"big cat" cat  -sports
-"all sports" -golf
-"professional golfers"
-"pga golfers" -"woods, tiger"]
```

   (b) **Detección y caracterización de nuevos sentidos**. Un ruido d se clasificará como nuevo sentido si verifica simultáneamente los siguientes criterios:

   Criterio 1. d se recupera en todas las consultas q_i. esta condición la cumplen los dos ruidos

```
sports/all sports/golf/
 professional golfers/
 pga golfers/woods, tiger

entertainment/movies/
 movies by genre/drama/
 epics/crouching tiger
```

   Criterio 2. d se asocia al menos al 10% de los primeros 30 documentos recuperados para cada consulta. Esta condición restringe el numero de nuevas acepciones de w a las cuales el algoritmo considera nuevos sentidos, admitiendo solamente aquellas con elevada presencia en las paginas web, y que por tanto son más proclives a producir altas cotas de ruido en búsquedas para otros sentidos.
   El ruido

```
entertainment/movies/
movies by genre/drama/
epics/crouching tiger
```

   no cumple esta condición puesto que para una de las consultas aparece en una sola ocasión
   Criterio 3. Un gran número de nuevos sentidos relevantes son nombres propios, referidos a compañías, productos comerciales, personajes de ficción etc. Este criterio trata de determinar si los ruidos que verifican los criterios precedentes se engloban dentro de éstos. Para verificarlo se realizan tres consultas:

   - Se forma una consulta con la palabra w, en minúsculas como término obligatorio, los términos de l_d, y todos los sinónimos de todos los sentidos conocidos como términos negativos. Llamamos P_1 al porcentaje de documentos recuperados que pertenecen al directorio de d. Por ejemplo, para

```
[+tiger sports "all sports"
 golf "professional golfers"
 "pga golfers" "woods tiger"
 -"Panthera tigris"]
```

     P_1 tiene un valor $30/30 = 1$.
   - Se forma una consulta idéntica, con la palabra w en mayúsculas.

Se calcula un porcentaje análogo P_2. Para el ejemplo anterior, el valor de P_2 es $30/30 = 1$.

- Se forma una tercera consulta, similar a las otras, ahora con la palabra w obligatoria en minúsculas y negativa en mayúsculas. Se le asigna también un porcentaje de documentos P_3. En nuestro ejemplo, $P_3 = 1/30 = 0.33$.

El criterio se cumple cuando

$$P_3 < P_1 \leq P_2$$

Es decir, si la precisión asociada a d decrece cuando se excluye el término en mayúsculas, y aumenta (o al menos no disminuye) si la palabra se considera exclusivamente en mayúsculas (señalando un nombre propio). En el ejemplo, $0.33 < 1 <= 1$, así que esa condición se cumple.

Una vez que el nuevo sentido queda detectado, procedemos a caracterizarlo de dos maneras complementarias:

- El nuevo sentido se define como *one of/part of* < del termino previo en la lista de dominio l_d >
- d se asocia al nuevo sentido como información de dominio.

Por ejemplo, un nuevo sentido generado por el algoritmo es:

`tiger newsense1 (tiger woods)` is a part of/one of `pga golfers`.

y se le asigna la etiqueta de dominio:

```
[sports,"all sports", golf,
"professional golfers",
 "pga golfers"]
```

3 Evaluación

Hemos seleccionado dos conjuntos diferentes de nombres para probar el algoritmo: un conjunto susceptible de tener nuevos sentidos, por un lado, y un conjunto con el menor sesgo posible, por otro.

El primero de ellos es el conjunto de hipónimos de *cat 6* en WordNet (*any of several large cats typically able to roar and living in the wild*). Los nombres de animales, bastante profundos en la jerarquía de WordNet, parecen buenos candidatos a tener nuevos sentidos y ruido asociado en consultas sobre Internet. Los nombres bajo *cat 6* son *cheetah, jaguar, leopard, lion, panther, tiger.*

Como conjunto no sesgado hemos tomado una selección de nombres utilizados en la primera competición SENSEVAL de sistemas de resolución de la ambigüedad léxica. Hemos considerado los nombres asociados adecuadamente a WordNet, un total de 17. Este conjunto es una referencia más estándar, formada por nombres con distintos tipos y grados de polisemia, no necesariamente oportunos para la aplicación de nuestro modelo.

3.1 Extracción de etiquetas de dominio

La Tabla 1 resume los resultados de la extracción de categorías de Altavista, y su clasificación en a) etiquetas de dominio para algún sentido ya existente, b) etiquetas de ruido que no derivan en nuevos sentidos, y c) etiquetas de ruido que indican un nuevo sentido relevante. En este apartado nos centramos en la primera columna de esa tabla.

Para el conjunto de hipónimos de *cat 6* el sistema predice 6 etiquetas de dominio para un total de 6 nombres (1 etiqueta por nombre), con un acierto del 100%. Entre las asignaciones correctas tenemos, por ejemplo:

```
jaguar#1 <-
[library,sciences,animals & wildlife,
  mammals,wildcats]

leopard#2 <-
[library,sciences,animals & wildlife,
  mammals,wildcats]
```

Para el conjunto de nombres procedente de SENSEVAL, el algoritmo propone un total de 23 etiquetas (1.35 etiquetas por palabra en promedio), 22 de las cuales son correctas. Algunos ejemplos son:

```
onion #3 <-
[lifestyle,gardening,plants,vegetables,
  varieties]
```

(onion 3 = *pungent bulb, kind of vegetable*)

```
giant #7 <-
[library,sciences,astronomy & space,
 celestial bodies,stars,
 stellar evolution]
```

	# Etiquetas	# correctas	Ruido	# correctas	# nuevos sentidos	# correctos
cheetah	1	1	-	-	1	1
jaguar	1	1	1	1	1	1
leopard	1	1	1	1	-	-
lion	1	1	4	4	-	-
panther	-	-	3	3	2	2
tiger	2	2	1	1	1	1
Total	**6**	**6(100%)**	**10**	**10**	**5**	**5**
accident	1	0	0	0	-	-
band	14	14	6	6	-	-
behavior	-	-	-	-	-	-
bet	-	-	-	-	-	-
bitter	-	-	-	-	1	1
excess	-	-	1	1	-	-
fascination	-	-	-	-	-	-
float	-	-	0	0	-	-
giant	1	1	7	7	-	-
hurdle	-	-	-	-	-	-
onion	1	1	-	-	-	-
promise	-	-	-	-	1	1
rabbit	2	2	1	1	-	-
sack	0	0	1	1	-	-
sanction	-	-	-	-	-	-
shake	1	1	2	2	-	-
shirt	3	3	2	2	-	-
Total	**23**	**22(96%)**	**20**	**20**	**2**	**2**

Tabla 1: Extracción de etiquetas de dominio y nuevos sentidos.

(giant 7 = *a very bright star of large diameter and low density*)

```
shirt #1 <-
[lifestyle,fashion,men's & women's wear,
women's wear, casual clothes]
```

(shirt 1 = *a garment worn on the upper part of the body*)

Un caso peculiar es *band 5*, que tiene 14 contextos correctos asignados (más de la mitad de las etiquetas obtenidas). *band 5* se refiere a bandas de música, y se encuentran etiquetas de este tipo:

```
band#5 <-
[entertainment,music,genres,alternative
& rock,classic rock, artists]

band#5 <-
[entertainment,music,genres,pop,disco,
   artists}

band#5 <-
[entertainment,music,genres,jazz,forms
& styles,big band & swing,artists,
   other artists]
```

Esto indica que debemos buscar un método para colapsar etiquetas redundantes o, alternativamente, utilizar las etiquetas para crear sentidos más especializados de uno determinado. En el caso de *band*, las etiquetas anteriores pueden producir las especializaciones *alternative & rock band, disco band, jazz band*, tomando como rasgo el subdominio donde las tres etiquetas divergen.

Es necesario recalcar que, al aplicar el algoritmo, sólo se consideran aquellos nombres que devuelven al menos 30 documentos para cada una de las consultas realizadas. Consideramos que, en caso contrario, la información que se obtiene no es suficientemente fiable. Algunas palabras, como *behaviour* o *hurdle*, no superan este valor de corte.

3.2 Detección de nuevos sentidos

La Tabla 1 muestra también las predicciones de ruidos y nuevos sentidos. Evidentemente, la detección de contextos esta marcada por la cobertura, puesto que ninguna etiqueta identificada como ruido/nuevo sentido admite una clasificación manual como contexto valido para un sentido ya existente. Para los dos conjuntos de prueba, las clasificacio-

Sentidos	%**documentos correctos consulta orig.**	%**documentos correctos +etiqueta 1**	**porcentaje absoluto de mejora**
band 1	3%	100%	+97
band 5	100%	100%	=
band 8	3%	70 %	+67
giant 7	20%	70%	+50
onion 3	27%	50%	+23
rabbit 1	27%	100%	+73
shake 3	33%	50%	+17
shirt 1	100%	100%	=
Total	**39%**	**80%**	**+41**
cheetah 1	43%	70%	+27
jaguar 1	20%	60%	+40
leopard 2	23%	93%	+70
lion 1	23%	97%	+74
tiger 2	43%	100%	+57
Total	**31%**	**84%**	**+53**

Tabla 2: Refinamiento de consultas con información de dominio

nes como ruido o nuevo sentido se han considerado manualmente un 100% correctas. La distinción entre "ruido"y "nuevo sentido"se hace con respecto a la densidad de aparición en Internet, de forma que no admite una evaluación manual directa más allá del criterio automático. Asumimos entonces que una etiqueta clasificada como "ruido"es correcta si manualmente no la encontramos relacionada con un sentido ya existente.

Dado el carácter especialmente dinámico de la recuperación de información procedente de Internet a través de sistemas de búsqueda como Altavista, los resultados obtenidos pueden variar en cortos espacios de tiempo. Si bien hemos constatado esta inestabilidad, las cifras apenas sufren variaciones globales.

No resulta sorprendente comprobar que el algoritmo es más productivo en el caso del conjunto de prueba dado por *cat*, en el que se predicen 5 nuevos sentidos para 6 nombres:

```
cheetah wheelies <-
 [entertainment,music,genres,alternative
  & rock, rock & roll,artists s]

jaguar <-
 [shopping,automotive,makes, models
  & clubs, makes h-l]

pink panther <-
 [entertainment,movies,movies by title,
  titles p]

panther <-
 [shopping,automotive,motorcycles,
  makes & models]

woods, tiger <-
[sports,all sports,golf,professional
 golfers, pga golfers]
```

Para el conjunto derivado de SENSEVAL, el sistema produce 2 nuevos sentidos (0.12 sentidos por palabra), clasificados manualmente como correctos.

3.3 Refinamiento de consultas y obtención de corpora

La Tabla 2 muestra los resultados obtenidos al refinar las consultas originales con la información aportada por las etiquetas de dominio obtenidas. Las entradas de la tabla corresponden a todos los sentidos de WordNet para los que se ha localizado una etiqueta de dominio. Cuando se ha obtenido más de una etiqueta, utilizamos la primera recuperada por el sistema. Para cada consulta, se han verificado manualmente los treinta primeros documentos que devuelve Altavista.

La primera columna refleja el porcentaje

de documentos en los que las ocurrencias de la palabra se corresponden con el sentido buscado con la consulta original (tal y como se formula en la sección 2.1). La segunda columna contiene la misma información, pero para la consulta refinada que incluye los términos de la etiqueta de dominio asignada por el algoritmo. La tercera columna muestra la mejora porcentual absoluta al pasar de la consulta original a la consulta refinada.

Puede apreciarse que la precisión crece, en promedio, del 39% al 80% para los nombres de SENSEVAL, y del 31% al 84% para los hipónimos de *cat 6*. Estos resultados sugieren firmemente que la etiquetación de sentidos utilizando categorías de Internet puede ser muy útil para producir corpora basados en la red.

4 Conclusiones

Hemos expuesto cómo un algoritmo relativamente simple y directo puede producir información útil para extender un diccionario con nuevos sentidos relevantes para llevar a cabo búsquedas en Internet y, simultáneamente, para dotar de etiquetas ricas en información de dominio a los sentidos ya existentes. Hemos comprobado también como está información se puede usar de forma directa para mejorar la obtención de corpora anotados semánticamente a partir de Internet.

Nuestro trabajo en curso consiste en

- Diseñar estrategias para colapsar conjuntos de etiquetas asignadas a un mismo sentido, como en el caso de *band 5*.
- Comprobar empíricamente si los sentidos que reciben una misma etiqueta pueden o no agruparse.
- Realizar una evaluación más exhaustiva del sistema, incluyendo medidas de cobertura: ¿cuántos sentidos de dominio específico tienen una categoría Altavista que les resultaría adecuada? Y entre estos, ¿cuántos casos detecta nuestro algoritmo?
- Una vez refinado, aplicar el algoritmo sobre el conjunto de nombres en WordNet 1.6 y medir el impacto de la información obtenida sobre una tarea concreta de búsqueda de información multilingüe.

Agradecimientos

Este trabajo ha sido financiado parcialmente por la Comisión Interministerial de Ciencia y Tecnología, proyecto *Hermes* (TIC2000-0335-C03-01).

Referencias

[1] E. Agirre, O. Ansa, E. Hovy, and D. Martínez. Enriching very large ontologies using the www. In *Proceedings of the Ontology Learning Workshop*, 2000.

[2] E. Agirre and D. Martínez. Exploring automatic word sense disambiguation with decision lists and the web. In *Proceedings of the COLING Workshop on Semantic Annotation And Intelligent Content*, 2000.

[3] http://www.altavista.com.

[4] J. Gonzalo, A. Peñas, and F. Verdejo. Lexical ambiguity and information retrieval revisited. In *Proceedings of EMNLP/VLC'99 Conference*, 1999.

[5] R. Magnini and R. Prevete. Exploiting lexical expansions and boolean compositions for web querying. In J. Klavans and J. Gonzalo, editors, *Proceedings of the ACL-2000 workshop on Recent Advances in Natural Language Processing and Information Retrieval*, 2000.

[6] R. Mihalcea and D. Moldovan. An automatic method for generating sense tagged corpora. In *Proceedings of AAAI '99*, 1999.

[7] R. Mihalcea and D. Moldovan. Autoasc - a system for automatic acquisition of sense tagged corpora. *International Journal of Pattern Recognition and Artificial Intelligence*, 2000.

[8] G. Miller. Special issue, Wordnet: An online lexical database. *International Journal of Lexicography*, 3(4), 1990.