Esta tesis doctoral contiene un índice que enlaza a cada uno de los capítulos de la misma.
Existen asimismo botones de retorno al índice al principio y final de cada uno de los capítulos.

Ir directamente al índice

Para una correcta visualización del texto es necesaria la versión de Adobe Acrobat Reader 7.0 o posteriores

Aquesta tesi doctoral conté un índex que enllaça a cadascun dels capítols.
Existeixen així mateix botons de retorn a l'index al principi i final de cadascun deis capítols .

Anar directament a l'index

Per a una correcta visualització del text és necessària la versió d' Adobe Acrobat Reader 7.0 o posteriors.

Universitat d'Alacant
Universidad de Alicante

Desambiguación léxica mediante Marcas de Especificidad

Tesis Doctoral

Autor: **Andrés Montoyo Guijarro**

Directores:
Dr. Manuel Palomar Sanz
Dr. German Rigau Claramunt

Dèpto. de Lenguajes y Sistemas Informáticos
Universidad de Alicante

Alicante, 14 de mayo de 2002

A Susana y Jorge

Agradecimientos

Quisiera dar mi más sincero agradecimiento a todas aquellas personas que han animado, apoyado, contribuido y colaborado en la consecución de esta Tesis.

En primer lugar quiero destacar todo el apoyo, ayuda y soporte por parte de mis dos directores, Manuel Palomar y German Rigau. Ambos con inagotable paciencia y dedicación me han proporcionado todo tipo de ideas, consejos y revisiones en cada una de las tareas para la elaboración de esta Tesis.

También estoy muy agradecido a todos mis compañeros del Grupo de Procesamiento del Lenguaje y Sistemas Informáticos de la Universidad de Alicante, en especial a Rafa Muñoz y Armando Suárez, por el apoyo y ánimo incondicional recibido y sin el cual hubiera sido difícil la finalización del presente trabajo. Y a todos mis compañeros del Departamento de Lenguajes y Sistemas Informáticos que me han facilitado la realización de esta investigación.

En el capítulo personal, quiero hacer un agradecimiento con amor a Susana, mi mujer, que siempre me ha alentado y animado para que pudiera finalizar con éxito esta Tesis. Ella ha sido la más sacrificada durante el largo periodo dedicado al desarrollo de esta Tesis. Y a mi madre y hermanas por todo lo que me han dado y ayudado en el aspecto personal y profesional.

Alicante, Mayo de 2002

Andrés Montoyo

El tribunal encargado de juzgar esta tesis doctoral está constituido por los siguientes doctores:

Dr. Lluís Padró (Universidad Politécnica de Cataluña)
Dr. Eneko Agirre (Universidad del País Vasco)
Dr. Julio Gonzalo (U.N.E.D.)
Dr. Patricio Martínez (Universidad de Alicante)
Dr. Ferrán Pla (Universidad Politécnica de Valencia)

Índice General

Índice de Tablas

Índice de Figuras

1. Introducción

1.1 Procesamiento del Lenguaje Natural

La aplicación de las nuevas tecnologías a los sistemas de información actuales ha provocado una revolución que está cambiando a la gente su forma de trabajar, de comunicarse con los demás, de comprar cosas, de usar los servicios e incluso en el modo de como se educan y se entretienen. Uno de los resultados de dicha revolución es que se está incrementando el uso y tratamiento de grandes cantidades de información con un formato más natural para los usuarios que los utilizados por los antiguos formatos típicos de las computadoras. Es decir, se están incrementando las actividades que utilizan y tratan el lenguaje natural como por ejemplo la redacción y corrección de documentos, consultas a distancia de fuentes de información, traducción automática, uso de diccionarios y enciclopedias, etc.

Por todos estos motivos, las investigaciones en la comprensión y en el uso de forma automática de los lenguajes naturales se han incrementado considerablemente en los últimos años. En esta área, denominada Procesamiento del Lenguaje Natural (PLN), se estudian los diferentes problemas que genera el lenguaje en su tratamiento automático, tanto en conversaciones habladas como escritas.

Para diseñar un sistema de PLN se requiere conocimiento abundante sobre las estructuras del lenguaje, como son el morfológico, sintáctico, semántico y pragmático.

- El conocimiento morfológico proporciona las herramientas para formar palabras, es decir cómo las palabras son construidas a partir de unidades más pequeñas.

- El conocimiento sintáctico proporciona cómo se deben combinar las palabras para formar oraciones correctas, además de estudiar cómo se relacionan unas con otras.
- El conocimiento semántico proporciona qué significan las palabras y cómo estas se combinan para formar el significado completo de una oración.
- El conocimiento pragmático ayuda a interpretar la oración completa dentro de su contexto, es decir proporciona cómo el contexto afecta a la interpretación de las oraciones.

Todas estas formas de conocimiento lingüístico tienen asociado un problema común difícil de resolver, sus diferentes ambigüedades. Por este motivo, cuando se diseña un sistema de PLN, uno de los objetivos fundamentales es resolver sus múltiples ambigüedades (estructural, léxica, ámbito de cuantificación, función contextual y referencial) mediante la definición de procedimientos específicos para cada una de estas.

1.1.1 Ambigüedad léxica

En concreto en esta Tesis nos centraremos en la resolución de la ambigüedad léxica, la cual se presenta cuando, al asociar a cada una de las palabras del texto la información léxico-morfológica, hay palabras que tienen más de un sentido o significado. Se distinguen dos tipos de ambigüedad léxica:

1. La *ambigüedad léxica categorial* se presenta cuando una palabra aparte de tener diferentes significados, éstos pueden desempeñar diferentes categorías sintácticas en la oración. Como por ejemplo, la palabra cura que puede ser un *nombre* en la oración: "El cura bendijo los alimentos", y un *verbo* en la oración: "El médico cura al paciente en el hospital".
   Algunas de las técnicas usadas en la literatura para resolver este fenómeno se basan en *aproximaciones lingüísticas*, como por ejemplo, el sistema EnCG (Votilainen, 1988; Votilainen y Järvinen, 1995) basado en gramáticas de restricciones, y otras se fundamentan en *aproximaciones basadas en aprendizaje automático* como son los trabajos desarrollados para el español

en Padró (1998) donde combina n-gramas y restricciones manuales, Márquez (1999) donde combina distintas técnicas de aprendizaje automático como árboles de decisión, Pla (2000) y Pla *et al.* (2000), donde utilizan técnicas basadas en *n-gramas*.

2. La *ambigüedad léxica pura* se presenta en aquellas palabras que en función del contexto pueden tener un sentido u otro. De forma más precisa, se puede decir que la ambigüedad léxica pura puede referirse tanto a *homonimia* como a *polisemia*.

   - **Polisemia:** Se encuentra polisemia cuando a una palabra le corresponden, según el contexto, varios significados. Es una nueva demostración del principio de economía que rige las lenguas, pues en caso de monosemia, se necesitaría un número mucho mayor de palabras. Para deshacer la ambigüedad hay que apoyarse en el contexto, el cual nos indica cuál es el significado pertinente, que puede tener un sentido irónico o humorístico. Son ejemplo de polisemia:

     El cura bendijo los alimentos
     El médico realizó la cura a los enfermos
     Antonio es un inocente
     Es un alumno inquieto
     A tus amigos les gusta jugar

     En estos ejemplos hay palabras que pueden entenderse de muy diversos modos, según el contexto:

     cura: sacerdote, curación, cuidado
     inocente: ingenuo, puro, inofensivo, sencillo, cándido, tonto, simple, bobo
     jugar: apostar, arriesgar, divertirse, entretenerse, esparcirse, recrearse, funcionar, etc.

   - **Homonimia:** Como la polisemia, la homonimia ofrece varios sentidos para una sola palabra. Pero esto viene motivado por la evolución histórica de una lengua, que, con el paso del tiempo, va confundiendo diferentes palabras en una

única forma por evolución fonética. Por ejemplo, la palabra **bala** pueden tener varios sentidos:

bala como munición o
bala como paquete grande de algo.

Las palabras que presentan homonimia se dividen en dos tipos:

- **Homófonos:** se pronuncian igual, pero se escriben de forma diferente, porque alguno de sus grafemas se corresponden con el mismo fonema, o porque no se corresponden con ninguno. Por ejemplo, las palabras:

tubo	tuvo
onda	honda
ojear	hojear

- **Homógrafos:** se escriben y pronuncian igual. Es necesario acudir al artículo, al plural o al contexto para saber su significado. Por ejemplo, las palabras:

el corte	la corte
esposa	esposas

El trabajo que aquí se presenta se centra en la resolución de la ambigüedad léxica, y en concreto, en la resolución de la ambigüedad léxica pura pero sin tratar la homonimia de homófonos. Ya que este tipo de ambigüedad es un obstáculo en aquellas aplicaciones que precisan conocer el significado semántico, por tal motivo, la resolución de esta ambigüedad mejora la calidad de algunos campos de la investigación. La resolución de los distintos tipos de ambigüedades, comentadas anteriormente, es tarea del Procesamiento del Lenguaje Natural, pero en particular, el campo que se encarga del estudio y resolución del problema de la ambigüedad léxica pura se conoce como *Desambiguación del sentido de las palabras* (WSD, en inglés *Word Sense Disambiguation*).

1.2 Desambiguación del Sentido de las Palabras

En términos generales, la desambiguación del sentido de las palabras consiste en asociar una palabra dada de un texto con una definición de un sentido o significado, lo que permite distinguirla de otros significados atribuibles a esa palabra. Entrando más en detalle, WSD consistiría en preprocesar un texto no restringido con el fin de extraer un conjunto de características (pistas o indicios), para posteriormente usarlas para asignar a cada palabra del texto el sentido más probable, adecuado y eficiente. La mayoría de los sistemas de WSD, para solucionar este problema, lo que hacen es determinar los diferentes sentidos de cada palabra del texto de entrada utilizando una lista cerrada de sentidos (como los que hay en un diccionario), un grupo de categorías (como las de un tesauro) o un diccionario multilingüe para traducirla a otro lenguaje. Para posteriormente mediante el uso del contexto de la palabra a ser desambiguada asignar un sentido apropiado. Para realizar la asignación del sentido a cada palabra, se utilizan dos recursos de información:

1. *El contexto* de la palabra a ser desambiguada, el cual se obtiene con la información contenida dentro del texto en el que la palabra aparece, junto con la información lingüística sobre el texto, como la colocación, etc.
2. *Recursos de conocimiento externo* como son los recursos léxicos, enciclopédicos, así como recursos de conocimiento desarrollados manualmente, que proporcionan datos valiosos para asociar palabras con sentidos. La gran mayoría de técnicas utilizadas para solucionar el sentido de las palabras se aplican individualmente pero según McRoy (1992) se deberían combinar varias técnicas y recursos de información.

Antes de continuar hay que hacer una aclaración muy importante sobre la preocupación en la forma de comparar los trabajos realizados en la desambiguación del sentido de las palabras debido a la dificultad para definir un sentido. Pero sin embargo, sí que ha

habido un acuerdo general para que el problema de desambiguación morfo-sintáctica y desambiguación de los sentidos se puedan separar (Kelly y Stone, 1975). Por lo tanto, para los homógrafos (palabras de distinta significación que se escriben de igual manera; por ejemplo: cura(nombre) y cura(verbo), la desambiguación morfo-sintáctica logra la desambiguación del sentido de las palabras (no entre "el cura" y "la cura"). Por tal motivo el trabajo de desambiguar el sentido de las palabras se ha enfocado mayoritariamente en distinguir sentidos homógrafos que pertenecen a la misma categoría sintáctica.

A continuación se realizará una descripción de cómo han evolucionado a lo largo del tiempo los estudios sobre la desambiguación del sentido de las palabras. Así, uno de los puntos centrales del tratado de Bar-Hillel (1960) fue la dificultad que conlleva la desambiguación del sentido de las palabras. Por ejemplo, él afirmó que no veía cómo el sentido de la palabra *pen* pudiera desambiguarse automáticamente a partir de la oración siguiente: *The box is in the pen*, ya que esta oración tiene varios significados correctos dependiendo de los sentidos escogidos. Este argumento de Bar-hillel fue la base preliminar del informe Alpac (1966), el cual se considera como la causa directa del abandono de la gran mayoría de las investigaciones de traducción automática sobre los años sesenta.

Durante esta misma época, hubo un progreso considerable en el área de la representación del conocimiento, sobre todo en las redes semánticas (Masterman, 1962; Richens, 1996), las cuales se aplicaron inmediatamente a la desambiguación del sentido de las palabras. Se intentó hacer una "interlingua", basada en los principios lógicos y matemáticos para resolver el problema de la desambiguación, relacionando palabras de cualquier lenguaje a una representación semántica/conceptual común. De tal forma que la primera base de conocimiento implementada automáticamente fue construida desde Roget´s Thesaurus (Masterman, 1957).

En los últimos diez años, los estudios y trabajos para automatizar la desambiguación del sentido de las palabras han aumentado considerablemente, así como los trabajos que tienen una relación directa con la lingüística computacional, debido a la po-

sibilidad de lectura de textos electrónicos y desarrollo de métodos estadísticos para identificar irregularidades sobre los datos. Como consecuencia de lo expuesto anteriormente, hoy en día la desambiguación del sentido de las palabras es uno de los problemas más importantes en la investigación sobre el procesamiento del lenguaje natural. Sin embargo, hay que considerar que los últimos métodos para la desambiguación del sentido de las palabras no han contribuido a dar solución a todas las investigaciones hechas al problema de WSD anteriormente. De hecho, no han aparecido datos cuantitativos que presenten mejoras empíricas a través de la desambiguación del sentido de las palabras, excepto en los campos de Traducción Automática (Brown et al., 1991) y en Recuperación de Información (Fukumoto y Suzuki, 1996). Una razón del fracaso en la mejora de los métodos existentes es debido a la falta del conocimiento del problema en la investigación de la desambiguación del sentido de las palabras, lo cual nos estimula a investigar intensamente esta área de investigación.

En el trabajo de Wilks and Stevenson (1996) se afirma que la Desambiguación del sentido de las palabras es una "tarea intermedia" que ayuda, en gran medida, a otras tareas del procesamiento del lenguaje natural, así como cuando necesitamos comprender el lenguaje en otros campos de la investigación como:

- *Recuperación de Información (RI).* Estos sistemas padecen los efectos del ruido en aquellas palabras polisémicas, por lo tanto se recuperarán documentos que tienen las mismas palabras pero diferentes sentidos. Los resultados de varios experimentos (Krovets y Croft, 1992; Sanderson, 1994; Schütze, 1995; Fukumoto y Suzuki, 1996; Krovets, 1997; Schütze, 1998; Gonzalo et al., 1998, 1999) demuestran que la tarea de desambiguar el sentido de las palabras es fundamental para mejorar los sistemas de recuperación de información. En estos casos el sistema encuentra solo aquellos documentos que usan palabras con el sentido apropiado. Evidentemente, el sistema eliminará todos aquellos documentos que usan palabras con sentidos inapropiados. Por ejemplo, cuando se buscan documentos que tengan referencias judiciales, es deseable eliminar los documentos que contienen la

palabra "corte" asociada con el sentido de "realeza y confección", y recuperar los que tengan sentido de "ley". Otros investigadores como Krovets and Croft (1992) proponen indexar documentos por el sentido de las palabras mediante la utilización de un diccionario automático y Voorhees (1993) mediante la utilización de WordNet.

- *Traducción automática (MT).* La desambiguación del sentido de las palabras es fundamental para la traducción, porque una palabra puede tener múltiples traducciones en el lenguaje destino, y cada una de estas pueden estar asociadas a sentidos diferentes. Por ejemplo la palabra francesa *grille*, dependiendo del contexto, puede ser traducido al español como: *barandilla, verja, bar, mostrador, rejilla, escala, etc* (Weaver, 1955; Yngve, 1955). Otro ejemplo sería la palabra inglesa *duty*, la cual puede tener los siguientes sentidos *Tax* y *obligation*, y que se corresponden a las palabras en español *impuesto* y *obligación/deber* respectivamente (Brown et al., 1991; Dagan et al., 1994).
  Los primeros intentos en la desambiguación automática del sentido de las palabras se hicieron en el contexto de la traducción automática. En el trabajo de Weaver (1955) se afirma que si se examinan una a una las palabras de un texto es imposible determinar el significado de esas palabras. Pero sin embargo, si se examina esa palabra junto con otras (a izquierda y derecha) sí que se puede decidir su significado. Pero con esto surge una duda en cuanto a cual es el valor mínimo de palabras a izquierda y derecha para obtener el significado correcto de la palabra. Un experimento, anterior al memorandum de Weaver (1955), hecho por Kaplan (1950) contesta en parte a la cuestión anterior. Este observó que la solución del sentido, dadas dos palabras a izquierda y derecha de la palabra a ser desambiguada, no era ni mejor ni peor que cuando se daba la oración completa. Otros investigadores como (Masterman, 1962; Koutsoudas y Korfhage, 1956; Gougenheim y Michéa, 1961; Choueka y Lusignan, 1985) también han descrito la misma experiencia.
  En un principio la TA se dirigió hacia los textos técnicos y hacia textos a partir de dominios particulares. Weaver (1955) discute el papel del dominio en la desambiguación del sentido, diciendo

que dentro de un contexto general de un artículo matemático cada palabra tiene un único significado. Posteriormente otros investigadores como Gale *et al.* (1992c) reiteraron estas palabras, por eso se desarrollaron muchos diccionarios especializados en los comienzos de la traducción automática.

El texto de Weaver (1955) también nos indicaba la utilidad de la estadística en el análisis del lenguaje comentando que los estudios semántico estadísticos deberían ser entendidos como un paso inicial completamente necesario. Varios trabajos de Harper (1957a; 1957b) siguieron esta aproximación en TA, haciendo estimaciones del grado de polisemia en textos y diccionarios.

- *Análisis temático del contenido*[1]. Los autores Stone *et al.* (1966; 1969; 1975) y más recientemente Litowski (1997) han admitido la necesidad de desambiguar el sentido de las palabras para este tipo de análisis. Para ello, se analiza cómo se distribuyen las categorías predefinidas de las palabras a lo largo del texto, con el fin de obtener aquellas palabras que nos indican una idea, tema o concepto determinado. Y la necesidad de desambiguar el sentido de las palabras en estos análisis sirve para incluir solamente aquellas palabras que tengan sentido apropiado.
- *Análisis sintáctico.* A menudo, cuando las relaciones sintácticas se asocian con el contenido semántico, falla el análisis sintáctico al identificar la estructura sintáctica de una oración. Los sintagmas preposicionales y las estructuras predicado-argumento asociadas con las restricciones seleccionales son ejemplos de este tipo de problema. Por tal motivo, para solucionar el problema y obtener el significado correcto de las oraciones, se requiere el contenido semántico de las entradas léxicas. Hay varios métodos que resuelven estos dos tipos de ambigüedad (Lytinen, 1986; Maruyama, 1990; Ravin, 1990; Nagao, 1994). También, la desambiguación del sentido de las palabras se utiliza en la fase de análisis sintáctico para etiquetar correctamente las palabras. Por ejemplo, en la oración "La banca ha aumentado sus intereses" es necesario desambiguar el sentido de la palabra "banca", la cual puede significar "asiento" o "entidad financiera". El pri-

[1] Con Análisis temático del contenido nos referimos a la búsqueda del tópico del discurso.

mero está en masculino y el otro en femenino, por lo tanto hay que desambiguarlo para etiquetarlo correctamente como nombre femenino.

- *Tratamiento de textos.* Se ha despertado un gran interés con el tratamiento de los textos, debido a la generalización de la mensajería y edición electrónica, que ha originado un incremento de la información de forma textual. Para manejar y acceder de forma eficiente a la información disponible, es necesario contar con herramientas que faciliten sin manipulación el trabajo con esos textos. A consecuencia de este interés por el tratamiento de textos, la desambiguación del sentido de las palabras se utiliza para la corrección ortográfica, por ejemplo en las palabras con acento "perdida" (Que no tiene o no lleva destino determinado) y "pérdida" (Carencia, privación de lo que se poseía) (Yarowsky, 1994a,b). O en los casos en los que hay que reconstruir las mayúsculas de "EL ESCRIBE EN EL MARCA" a "El escribe en el Marca". Y también, se utiliza como ayuda en los editores de textos para sugerir palabras alternativas semánticamente equivalentes a las encontradas en un texto cuando se está editando el mismo.
- *Enfoques del PLN basados en clases semánticas.* Los planteamientos presentados en los trabajos de Ker, Resnik, Ribas y McCarthy (1997; 1993a; 1995b; 2001) se benefician de la desambiguación del sentido de las palabras, ya que relacionan cada palabra de entrada con una clase semántica (casi siempre a partir de una taxonomía de tesauro).
- *Sistemas de Dialogo y Extracción de Información (IE).* Los sistemas que extraen información necesitan obtener estructuras semánticas, y para ello muchas veces hay que conocer o considerar el significado de las palabras. En el trabajo de Kilgarriff (1997) se describe que los sistemas de Dialogo y Extracción de Información emplean normalmente la representación del conocimiento mediante el dominio específico en el que se aplica, en vez de usar la desambiguación del sentido de las palabras con objeto de solucionar la ambigüedad léxica.
- *Ventajas para la lexicografía.* En el trabajo Kilgarriff (1997) se presentan las ventajas que tiene la desambiguación del sentido

de las palabras en la lexicografía. Una de las ventajas presentadas es anotar el sentido a los datos lingüísticos, para reducir considerablemente el esfuerzo que tienen que hacer los lexicógrafos cuando deben clasificar grandes corpus en relación al uso de las palabras por sus distintos sentidos.

Con el objetivo de centrar el problema a resolver en esta Tesis diremos que el trabajo que aquí se presenta consiste en la resolución de la ambigüedad léxica pura (semántica) en textos de dominios no restringidos que tenga un repositorio de sentidos en una lengua concreta como inglés, español, italiano, etc. Además, consideramos y así demostramos a lo largo de este trabajo, que una resolución adecuada de la ambigüedad léxica pura puede mejorar considerablemente a otros campos del procesamiento del lenguaje natural, como por ejemplo los comentados anteriormente.

1.3 Objetivo

El objetivo principal de este trabajo es el estudio de un método para la resolución de la ambigüedad léxica pura en textos de dominios no restringidos. Para la consecución de este objetivo se plantean las siguientes líneas de actuación:

- Estudio de las ambigüedades léxicas y en concreto la ambigüedad léxica pura como aplicación del método para su resolución.
- Estudio y desarrollo de estrategias de resolución de la ambigüedad léxica pura. Para ello se tomarán como hipótesis inicial los trabajos previos de la resolución de la ambigüedad léxica.
- A partir de la clasificación de los distintos tipos de métodos para la resolución de la ambigüedad léxica pura, elegir y definir el ámbito de trabajo. En concreto se ha elegido el ámbito de trabajo de los métodos basados en el conocimiento. Porque estos métodos tienen la ventaja de no necesitar procesos de entrenamiento, ni codificación manual de las entradas, ni etiquetado manual.

- Estudio de la influencia de la estructura taxonómica del recurso léxico utilizado en la resolución de la ambigüedad léxica pura con el objetivo de adquirir conocimiento.
- Desarrollo de un proceso de ajuste del método que garantice la mejor estrategia de resolución de la ambigüedad léxica pura.
- Desarrollo de un proceso de evaluación del método con el fin de obtener resultados finales de efectividad en la resolución de la ambigüedad léxica pura.
- Estudio, definición y desarrollo de una aplicación e integración del método de Marcas de Especificidad orientado al enriquecimiento de WordNet mediante las categorías, por ejemplo, de los sistemas de clasificación de noticias.
- Desarrollo de un proceso de evaluación del nuevo sistema para enriquecer WordNet con categorías de los sistemas de clasificación para comprobar su efectividad.

Junto con estas líneas de actuación se tendrá en cuenta que el método queda abierto para plantear nuevas estrategias de resolución como por ejemplo la resolución léxica en verbos, adjetivos y adverbios. Además aunque el método de WSD se aplica al enriquecimiento de un recurso léxico, y en particular de una base de conocimiento léxica, su aplicación no debe quedar solo limitada a esto. Cualquier aplicación computacional que requiera la comprensión, interpretación y/o generación del lenguaje natural puede ser receptora de un módulo de resolución de la ambigüedad léxica pura para su mejor funcionamiento.

1.4 Organización de la Tesis

Esta Tesis se ha estructurado en seis capítulos:

- Este primer capítulo es una introducción al problema del procesamiento del lenguaje natural, en concreto al problema de la ambigüedad léxica pura, en el que se presenta la investigación que se encarga del estudio y resolución de este problema, así como las mejoras que aporta a otras áreas del procesamiento del lenguaje natural. También se describe en este capítulo la estructura de esta memoria.

- El segundo capítulo presenta una descripción de los conceptos más relevantes en Desambiguación del sentido de las palabras y se hace una profunda revisión bibliográfica general del estado actual en las investigaciones sobre WSD. Para ello este capítulo presenta una clasificación de los distintos métodos de desambiguación y finalmente se presentan los sistemas de WSD que participaron en las competiciones SENSEVAL-1 y SENSEVAL-2 con una breve descripción de los métodos más relevantes en dicha competición y todos los métodos basados en el conocimiento que se relacionan con el método presentado en este trabajo.
- En el tercer capítulo se presentan las principales aportaciones de nuestra investigación: en concreto, la propuesta del método para la resolución de la ambigüedad léxica de nombres en cualquier lengua que tenga un repositorio de sentidos organizados como una base de conocimiento (Método de Marcas de Especificidad). Para ello, en este capítulo presentamos inicialmente la arquitectura del sistema de Procesamiento del Lenguaje Natural (PLN) utilizado junto con los recursos y herramientas lingüísticas utilizadas por el método, para posteriormente presentar detalladamente el funcionamiento del método propuesto.
- El cuarto capítulo incluye la evaluación completa de este método propuesto, así como todo el proceso de ajuste llevado a cabo para obtener la mejor versión. Para ello este trabajo ha realizado inicialmente una fase de adaptación, con el objetivo de mejorar el método hasta conseguir una versión final, para posteriormente realizar la fase de evaluación, la cual consiste en presentar resultados de funcionamiento del método de Marcas de Especificidad. Tanto para la fase de ajuste/adaptación como para la de evaluación se presenta una introducción para ambos procesos, así como una descripción detallada del entorno experimental, mediante la presentación de los recursos empleados y las distintas ventanas contextuales empleadas en los experimentos.
- En el capítulo quinto se presenta la aplicación del método de Marcas de Especificidad a un método para enriquecer semánticamente el recurso léxico WordNet con categorías o etiquetas de otros sistemas de clasificación. El sistema de clasificación utilizado para etiquetar automáticamente los registros de

la base de conocimiento léxica ha sido IPTC Subject Reference System. Para llevar a cabo este trabajo se realiza inicialmente una introducción a los sistemas de clasificación y una presentación particular al sistema de clasificación IPTC. A continuación, se describe detalladamente el método propuesto para enriquecer semánticamente WordNet con categorías IPTC[2] (Versión IPTC/1), así como las características del diseño e implementación de la interfaz construida para extender y mejorar la base de datos léxica WordNet. Y finalmente, se muestran los experimentos realizados al método propuesto así como las conclusiones obtenidas al analizar sus resultados.

- El capítulo sexto muestra las conclusiones y beneficios de esta Tesis junto con una propuesta de trabajos futuros.
- Por último se presenta una relación bibliográfica de los trabajos que se han utilizado como referencia bibliográfica para el desarrollo de esta Tesis.

1.5 Resumen del enfoque del método de Marcas de Especificidad

En este trabajo se ha desarrollado un método para resolver la ambigüedad léxica pura (semántica) en textos de dominios no restringidos en cualquier lengua que tenga un repositorio de sentidos organizado como una base de conocimiento léxica. El método de resolución de la ambigüedad léxica pura propuesto se basa en el uso de conocimiento lingüístico (información léxica y morfológica) y de conocimiento a partir de las relaciones léxicas y semánticas de un recurso externo (taxonomía de nombres a partir de la base de conocimiento léxica utilizada), ambos independientes del dominio.

Gracias al uso de información no dependiente del dominio, este método de resolución de la ambigüedad léxica, al que hemos denominado Marcas de Especificidad, está preparado para ser aplicado a sistemas que trabajen sobre cualquier dominio y sobre cualquier lengua que tenga un repositorio de sentidos predefinido.

[2] `http://www.iptc.org`

2. Desambiguación del sentido de las palabras: Métodos de resolución

El objetivo de este capítulo es presentar una profunda revisión bibliográfica general del estado actual en las investigaciones sobre WSD. Para ello, en primer lugar, se describirán los conceptos más relevantes en WSD, junto a una primera clasificación de los distintos métodos de desambiguación: basados en corpus y basados en conocimiento. En segundo lugar, según la clasificación propuesta, se presentan los distintos métodos de desambiguación basados en corpus. A continuación, se presentan los métodos de desambiguación basados en conocimiento, así como los métodos híbridos, los cuales combinan varias fuentes de conocimiento y diferentes técnicas para explotar dicho conocimiento. Finalmente, se presenta una clasificación alternativa de los sistemas WSD, desde el punto de vista de si requieren ejemplos etiquetados manualmente o no.

2.1 Introducción

La tarea de WSD pertenece al campo del PLN que se encarga del estudio y resolución del problema de la ambigüedad léxica pura. Su tarea consiste en identificar el significado concreto con que aparece una palabra en un texto con un determinado contexto. Para desambiguar una palabra que aparece en un determinado lugar de un documento se utilizan dos fuentes de información: definiciones de un recurso léxico y el contexto de aparición.

- *Definiciones de un recurso lingüístico.* Una palabra puede tener asociados diferentes sentidos y cada sentido debe tener asociado una definición, la cual consiste en un conjunto de palabras.

Esta definición se puede obtener de un recurso lingüístico externo como puede ser un diccionario electrónico ó una base de conocimiento léxica.

- El *contexto* en el que aparece la palabra a desambiguar estará formado por las palabras que aparecen próximas a él en el texto del documento.

La desambiguación de una palabra en un documento, a partir de los dos recursos explicados anteriormente, consiste en asignar el sentido apropiado al contexto en el que aparece esa palabra.

Antes de continuar es importante realizar una aclaración sobre la dificultad que tiene la automatización del proceso WSD, debido a los problemas que conlleva como al grado de granularidad tan fina que proporcionan los diccionarios en la división de los sentidos, a que el diccionario no contenga el sentido apropiado, a que el sentido del diccionario se demasiado particular, a que puedan aplicarse varios sentidos, etc. Así, algunos autores como Slator y Wilks (1987) han afirmado que la división de los sentidos que proporcionan los diccionarios es normalmente muy fina para las tareas de PLN. Esto es un problema muy grave para la tarea de WSD. Y la consecuencia de esto es que se requiere realizar una elección de sentidos extremadamente dificultosa incluso para lexicógrafos expertos, debido a que distinción de los sentidos realizada en algunos diccionarios es difícil de hacer incluso por parte de los lexicógrafos expertos. Para solucionar el problema de la automatización de WSD se han propuesto distintos enfoques, pero el problema sigue sin solucionarse definitivamente.

Los métodos, que se utilizan para desambiguar el sentido de las palabras automáticamente, se pueden clasificar de distintas formas: supervisados, no-supervisados, basados en ejemplos de corpus, basados en bases de conocimiento, métodos mixtos, etc.

Los métodos supervisados requieren corpus semánticamente anotados de modo manual por un humano para posteriormente entrenar al sistema, además de un recurso léxico que suministra el sentido que se corresponde con la anotación hecha. Esta tarea manual de anotación semántica es muy laboriosa y costosa de realizar, ya que se necesitan conjuntos de entrenamiento de

gran tamaño. Sin embargo, hay situaciones en las que no se dispone de estos recursos para poder realizar la desambiguación, por lo que se debe hacer de forma no supervisada. Así, sistemas no-supervisados se definen como aquellos métodos que no necesitan estos ejemplos para realizar la desambiguación del sentido de las palabras, es decir obtienen esta información automáticamente.

Los métodos basados en ejemplos de corpus (en inglés, WSD Corpus-Based) utilizan ejemplos de usos de las palabras ("bancos de datos"), que previamente han sido anotados semánticamente, para la desambiguación del sentido de las palabras. A partir de esta información y aplicando alguna aproximación obtienen el sentido de las palabras, por lo tanto estos métodos dependen de la disponibilidad de los corpus anotados semánticamente y su dependencia de los datos utilizados en la fase de entrenamiento.

Los métodos basados en bases de conocimiento (en inglés, WSD Knowledge-Driven) se caracterizan porque su trabajo de desambiguación consiste en emparejar la palabra a ser desambiguada con cualquier información de un recurso de conocimiento externo (diccionario, base de conocimiento, etc). Estos métodos utilizan recursos de conocimiento léxico preexistentes, por lo que evitan la necesidad de utilizar grandes cantidades de información de entrenamiento para desambiguar el sentido de las palabras.

Los métodos mixtos se caracterizan porque combinan distintos recursos de conocimiento léxico y utilizan diferentes técnicas con el objetivo de desambiguar el sentido de las palabras. Estas técnicas empleadas se diferencian en el tipo de recurso léxico empleado en los distintos pasos del método, en las características de los recursos utilizados durante la desambiguación y en las medidas utilizadas para comparar similitudes entre unidades léxicas.

En esta tesis se ha creído conveniente utilizar como base la clasificación propuesta por (Ide y Véronis, 1998), aunque hay otros autores que realizan una clasificación diferente como (Wilks y Stevenson, 1996). A continuación se presentan las investigaciones de mayor relevancia en WSD según los criterios de clasificación expuestos anteriormente.

2.2 Métodos basados en corpus

Los métodos basados en corpus se caracterizan porque su trabajo de desambiguación consiste en emparejar la palabra a ser desambiguada con modelos obtenidos a partir de los contextos de ejemplos[1] de usos de esa palabra.

Estos métodos basados en corpus usan normalmente un corpus con las palabras de cada oración anotadas semánticamente, bien manualmente o automáticamente, con su sentido correcto. A los métodos que requieren anotación manual de los corpus se les llaman métodos *Supervisados*. A los que excluyen la tarea de supervisarlos se les llama métodos *No supervisados*. Los métodos supervisados requieren conjuntos de entrenamiento normalmente de gran tamaño con anotaciones manuales del sentido de las palabras. El desarrollo de estos corpus anotados semánticamente proporcionan el medio para inducir automáticamente reglas o modelos probabilísticos para la desambiguación. Sin embargo, la creación manual de corpus anotados semánticamente a nivel de sentido es una tarea difícil, tediosa y extremadamente costosa (Kilgarriff y Palmer, 2000).

A continuación se van a clasificar los métodos basados en corpus de acuerdo a su mecanismo de inducción. Es decir, al mecanismo que nos permite obtener modelos, leyes o principios a partir de los fenómenos, hechos o casos particulares. La clasificación sería: *métodos basados en reglas*, *métodos basados en modelos probabilísticos*, y *métodos basados en la similitud semántica.*

La clasificación anterior se complementará con otros tipos de métodos basados en corpus como *métodos basados en el uso de corpus bilingües* y *métodos basados en propiedades discursivas.* En las siguientes sub-secciones se describirán detalladamente cada una de las clasificaciones anteriores.

[1] Cada aparición de la palabra a ser desambiguada en el corpus se encuentra anotada con el sentido apropiado conforme al contexto donde aparece

2.2.1 Métodos basados en reglas

Los métodos basados en reglas utilizan reglas selectivas[2] asociadas con cada sentido de la palabra. Dada una entrada de una palabra polisémica, el sistema selecciona el sentido que cumpla satisfactoriamente las reglas que determinan a uno de los sentidos. Estos métodos tienen un problema asociado dependiendo de la granularidad asociada a las reglas selectivas, ya que sufren el riesgo de seleccionar incorrectamente el sentido tanto si las reglas son demasiado específicas como si son muy generales. Dentro de los métodos basados en reglas se puede hacer una subdivisión dependiendo del formato de las reglas.

Métodos basados en restricciones de selección. Las restricciones de selección limitan el sentido de una palabra dependiendo de las clases semánticas de las palabras con que aparece. Además, las restricciones de selección para ser operativas deberían aplicarse junto con la información sintáctica de las palabras. Así, los lexicones[3] que se aplican en los métodos deberían incluir tanto información de la posición sintáctica de las palabras como la información de tipo semántico que limitan o restringen. Por ejemplo, uno de los sentidos comunes del verbo "beber" restringe su sujeto a un rasgo "animal" y su objeto directo a un "líquido".

Por lo tanto, los métodos basados en las restricciones de selección aplican limitaciones a los argumentos de una palabra dada (normalmente interpretada como un predicado) y ayudan a seleccionar el sentido correcto tanto de los verbos como de las palabras que forman sus argumentos.

En las oraciones mostradas en el ejemplo 1, las restricciones de selección nos pueden ayudar a distinguir entre los dos sentidos (*"contratar"*/ *"utilizar"*) del verbo *"emplear"*.

(1) **1** *El jefe empleó nuevos trabajadores.*

[2] Se entiende por regla selectiva al conjunto de operaciones que deben llevarse a cabo para realizar una inferencia o deducción correcta.

[3] Un lexicón puede ser un diccionario electrónico o un tesauro. (Wilks et al., 1996)

2 *El jefe empleó su propuesta.*

En la oración ***1*** su sujeto se asocia con las características semánticas de *"humano/organización"* y su objeto directo con *"humano"*. Por lo tanto al tener a *"trabajadores"* como objeto directo, el sentido correcto para *"emplear"* es *"contratar"*. Sin embargo, en la oración ***2*** su sujeto se asocia con las características semánticas de *"humano/organización"* y su objeto directo con *"idea"*. Por lo tanto al tener a *"propuesta"* como objeto directo, el sentido correcto será *"utilizar"*.

Varios investigadores han aplicado las restricciones de selección para desambiguar el sentido de las palabras: los trabajos de Wilks (1975a) y Hirst (1987) proponen el uso de preferencias codificadas manualmente para resolver el sentido correcto de las palabras. Trabajos como el de Yarowsky (1993) o el de Ribas, Li & Abe y McCarthy (1995b; 1995; 2001) proponen el uso de conocimiento de la colocación de las palabras extraída de un corpus para decidir el sentido correcto en patrones semánticos predicado-argumento o adjetivo-nombre. Y el trabajo de Resnik (1993a) propone la identificación de las restricciones de selección como clases semánticas definidas en la taxonomía de WordNet (Miller et al., 1990), y solo aquellos nombres que están en la taxonomía y se han identificado sus clases cumplirán la restricción.

Métodos basados en árboles de decisión. Los árboles de decisión son una herramienta para tomar decisiones cuando se trata con información compleja. Los árboles de decisión proporcionan una estructura eficiente para establecer las distintas decisiones alternativas y evaluar lo que implica tomar esas decisiones. También es muy útil para clasificar información. Un ejemplo de árbol de decisión para aplicar el descuento a un cliente es el mostrado en la Figura 2.1. Cómo se puede ver la hoja marcada con "*" es la raíz del árbol y dependiendo si el cliente paga en un periodo superior o inferior a 10 días y de la cantidad a pagar se le aplicará un descuento.

Hay que destacar en este punto que los árboles de decisión se han aplicado muy poco a WSD, sin embargo, el algoritmo C4.5 (Quinlan, 1993) se ha utilizado como base de prueba para realizar

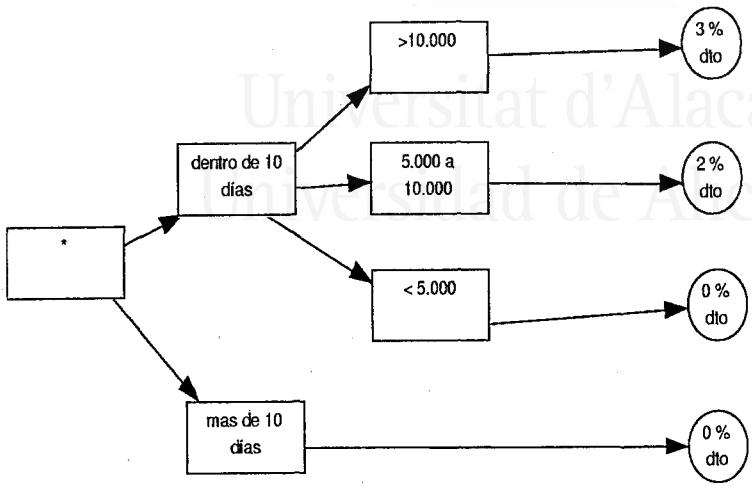

Figura 2.1. Árbol de decisión para aplicar descuento a cliente

la comparación de varios algoritmos que utilizan árboles de decisión. Las investigaciones de Mooney (1996) y de Pedersen (1997a) utilizaron el algoritmo C4.5 para comparar el rendimiento de varios métodos de WSD y Tanaka (1994) lo utilizó para adquirir reglas de traducción de verbos entre inglés y japonés. Como un verbo en inglés puede interpretarse como varios verbos diferentes en japonés, a esta última investigación se le puede considerar como un tipo de regla de inducción para WSD. Por ejemplo, el verbo en inglés *take* puede traducirse como *erabu (To choose)*, *tsureteiku (To take along)* o *motteiku (To take away)*. En la Figura 2.2 se muestra un fragmento del árbol de decisión para el verbo en inglés *take*.

Por ejemplo, si se tiene una oración con el verbo *take* seguido del objeto directo *him* y una preposición *to*, el árbol de decisión selecciona al verbo en japonés *tsureteiku (To take along)*, ya que en el árbol de decisión pasa por las ramas correspondientes a *object=him* y *prep=to*.

Esas reglas de traducción adquiridas sirven para aplicarlas a WSD, ya que dependiendo de las reglas que cumpla el verbo de

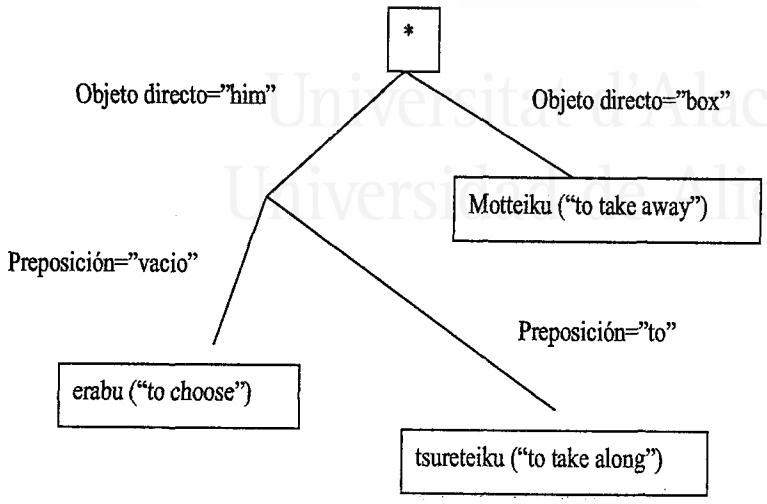

Figura 2.2. Un fragmento de árbol de decisión para el verbo en inglés *take*

entrada se obtendrá uno de los distintos sentidos posibles en japonés.

Métodos basados en listas de decisión. Las listas de decisión son estructuras para representar reglas y están compuestas de tuplas de la forma (condición, valor). Según Rivest (1987) las listas de decisión se pueden considerar como reglas *"si-entonces-sino"*, por lo tanto las condiciones excepcionales aparecen al principio en la lista mientras que las condiciones generales aparecen en los últimos lugares, y la última condición de la lista será aquella que acepte todos los casos. Es decir, cuando se hace una pregunta a la lista de decisión, esta actúa de la siguiente manera: cada condición en la lista de decisión se aplica secuencialmente hasta que se encuentre la condición que satisfaga a la pregunta realizada, y entonces, el valor que corresponda a esa condición se selecciona como contestación a la pregunta realizada. Los trabajos de investigación que han utilizado las listas de decisión en la WSD se comentan a continuación.

Los autores Agirre y Martinez (2000) realizaron un estudio detallado sobre el funcionamiento de las listas de decisión sobre dos

corpus públicos disponibles y un corpus adicional adquirido automáticamente a partir de la Web, mediante el uso de los sentidos de WordNet.

El trabajo de Yarowsky (1994b) propone la aplicación de las listas de decisión con el objetivo de resolver un tipo particular de ambigüedad léxica, la cual aparece cuando una palabra se asocia con múltiples pronunciaciones (debido a los acentos tanto del español como del francés). También, en otro trabajo, Yarowsky (1995) aplicó las listas de decisión a WSD. En los dos trabajos[4] anteriores de Yarowsky cada condición se corresponde a la colocación de una palabra en una oración y cada valor corresponde al sentido correcto de la palabra o a la pronunciación. Las colocaciones de las palabras las obtuvo de un gran corpus con el objeto de identificar automáticamente hechos o indicios efectivos (E), ya que la identificación manual de éstos es muy costosa. El grado de efectividad de estos hechos o indicios se calcula con la probabilidad de que un sentido dado esté más cercano que otro a estos hechos. Formalmente hablando, el grado de efectividad de estos hechos se calcula aplicando la Fórmula 2.1, la cual representa el ratio entre la probabilidad condicional de que el sentido s_1 y el sentido s_2 aparezcan con un hecho o indicio efectivo (E) dado. En la lista de decisión, cada hecho o indicio con sus sentidos posibles se clasifican en orden descendente respecto a su probabilidad.

$$\log \frac{P(s_1|E)}{P(s_2|E)} \tag{2.1}$$

A continuación, se expondrá un ejemplo de Yarowsky (1995) que utiliza un fragmento de una lista de decisión entrenada para desambiguar la palabra *plant* (*organism/factory*, en donde cada hecho o indicio indica la colocación de la palabra con su distancia asociada a *plant* o un patrón de su colocación específica. Dicho ejemplo, se muestra en la Tabla 2.1. Por lo tanto, si se da una entrada que contiene el patrón *plant height*, la interpretación que se le asocia a *plant* es *organism* y así sucesivamente para las demás entradas.

[4] En los trabajos de Yarowsky (1994b) y (1995), el número de sentidos candidatos para cada palabra se limitó a dos.

Hecho o indicio	Sentido	Probabilidad
Plant growth	Organism	10,12
Car (*within* $\pm$ *k words*)	Factory	9,68
Plant height	Organism	9,64
Union (*within* $\pm$ *k words*)	Factory	9,61
Equipment (*within* $\pm$ *k words*)	Factory	9,54
$\vdots$	$\vdots$	$\vdots$

Tabla 2.1. Un fragmento de lista de decisión para la palabra en inglés *plant*

El algoritmo de reglas de inducción llamado CN2 (Clark y Niblett, 1989) genera listas de decisión a partir de un conjunto de datos de entrenamiento previamente dado. En definitiva lo que hace es encontrar reglas y medir la importancia de esas reglas basándose en entropía y su cobertura sobre el rango de los datos de entrenamiento. La calidad de las listas de decisión depende en gran medida de la secuencia de las reglas, ya que reglas sencillas al final de la secuencia empeoran el proceso de desambiguación. Para resolver este problema en la última versión del algoritmo CN2 se utilizó una técnica denominada "reglas desordenadas" (Clark y Boswell, 1991). También, autores como Pedersen y Bruce (1997a) utilizaron el algoritmo de reglas de inducción CN2 en sus trabajos.

2.2.2 Métodos basados en modelos probabilísticos

La idea de los modelos probabilísticos consiste en: dada una palabra a desambiguar y un conjunto de palabras que la acompañan en la frase o párrafo (este conjunto de palabras se le denomina "contexto"), desambiguar el sentido de la primera utilizando información que aporta el contexto. Es decir, se obtiene un modelo probabilístico en función de un conjunto de hechos o indicios observados. Así, sobre la base de un hecho de un nuevo contexto y basándose en los hechos observados, se proporciona la probabilidad de un determinado sentido para este nuevo hecho. A continuación se detalla la idea intuitiva de probabilidad condicional, se expone algún ejemplo aclaratorio y se especifica su notación formal.

Formalmente, la probabilidad condicional de un evento A conociendo que el evento B ocurre ($P(B) > 0$) se calcula mediante la fórmula 2.2.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2.2}$$

La generalización de esta regla a múltiples eventos se usa en muchos métodos basados en modelos probabilísticos para desambiguar el sentido de las palabras (Leacock et al., 1993; Pedersen y Bruce, 1997b). Así, desde el punto de vista de la teoría de la probabilidad, la tarea de desambiguar el sentido de las palabras (WSD) consiste en seleccionar el sentido de una palabra que tenga la mayor probabilidad condicional para que a esa palabra le corresponda ese sentido. Para ello hace falta un corpus desambiguado para entrenar al método WSD, el cual es un conjunto de entrenamiento de ejemplos donde cada palabra ambigua que aparece en el corpus se anota con una etiqueta semántica. La notación que se usa a partir de ahora será la mostrada en la Tabla 2.2:

Símbolo	Significado
W	Palabra ambigua
$s_1, \cdots, s_k, \cdots, s_K$	Sentidos de la palabra ambigua w.
$c_1, \cdots, c_k, \cdots, c_K$	Contexto de w en el corpus.
$v_1, \cdots, v_k, \cdots, v_K$	Palabras usadas como características para desambiguar.

Tabla 2.2. Notación utilizada en esta sección

El teorema de Bayes, que se presenta para desambiguar el sentido de las palabras, consiste en observar aquellos rasgos o características (normalmente palabras) que están alrededor de una palabra ambigua en un contexto. Así cada rasgo aporta información muy útil sobre qué sentido de la palabra ambigua se usa comúnmente con ella. Formalmente, el teorema de Bayes se calcula con la fórmula 2.3.

$$P(s_k|c) = \frac{P(s_k) \times P(c|s_k)}{P(c)} \tag{2.3}$$

El valor de $P(s_k)$ es la "probabilidad previa" del sentido s_k, es decir, la probabilidad del s_k sin saber nada sobre el contexto. Pero $P(s_k)$ se actualiza con el factor $\frac{P(c|s_k)}{P(c)}$ el cual incorpora los hechos que se tienen en el contexto, obteniéndose como resultado final la "probabilidad posterior" $P(s_k|c)$. El valor de $P(c|s_k)$ se obtiene aplicando la fórmula 2.6.

Si lo que realmente se quiere es elegir la clase correcta entonces la fórmula 2.3 se puede simplificar. Esto se obtiene al eliminar $P(c)$, que es una constante para todos los sentidos y por lo tanto no influye en el su cálculo. Por lo tanto, con la fórmula 2.4 lo que se pretende es asignar la palabra w al sentido S'.

$$S' = \arg\max_{s_k} P(s_k|c) = \arg\max_{s_k} P(s_k) \times P(c|s_k) \qquad (2.4)$$

Generalmente, para estos métodos la entrada de datos se representa con un vector de características WSD con la siguiente estructura : $< F_1 = f_1, F_2 = f_2, \cdots, F_n = f_n >$, donde F_i y f_i son la característica número i y su valor respectivamente.

A continuación, se presentarán los métodos basados en modelos probabilísticos de acuerdo a: si primero se identifica un conjunto de características (rasgos) que pueden ser palabras, POS, palabras en ciertas posiciones sintácticas, etc. Y luego se calcula la probabilidad condicional de cada característica o rasgo respecto al sentido o clase.

- Primero se identifica un conjunto de características o rasgos, y en concreto diferentes conjuntos de características para cada una de las diferentes palabras polisémicas. Por ejemplo, aquellas palabras que aparecen junto con la palabra (*collocation words*) a desambiguar son palabras que típicamente se usan como características. Así, cada característica toma un valor binario, es decir un 1 cuando se da esa característica para los datos de entrada y un 0 cuando no.
  Los trabajos de Yarowsky (1992), Justeson y Katz (1995) proponen métodos que identifican automáticamente palabras con características informativas o relevantes. Particularmente, el trabajo de Yarowsky utilizó la "Información Mutua" entre la palabra w y el sentido s (para obtener los distintos sentidos de las

palabras utilizó las categorías semánticas definidas en el Thesauro Roget´s (Chapman, 1984)) para estimar el grado de característica informativa o relevante de la palabra w al sentido s. Intuitivamente hablando, la "Información Mutua" entre dos variables o fenómenos da un mayor valor cuando estas variables o fenómenos son más probables a aparecer conjuntamente. Es decir, mide la información común en las dos variables. Si dos palabras tienden a aparecer siempre juntas en el contexto, su relación debería ser fuerte. Pero si dos palabras nunca aparecen en el mismo contexto, su relación debería ser débil.
La "Información Mutua" de w y s se calcula aplicando la formula 2.5, siendo $P(w|s)$ la probabilidad de que w aparezca dando s (en el caso de Yarowsky, categoría Roget), y $P(w)$ la probabilidad de que w aparezca en el contexto[5]. Estos factores se estiman basándose en la distribución relativa de w y s en los datos de entrenamiento.

$$I(w;s) = \log \frac{P(w|s)}{P(w)} \tag{2.5}$$

La Tabla 2.3 muestra un ejemplo de palabras con características informativas relacionadas con las dos categorías Roget "ANIMAL" y "TOOLS" (Yarowsky, 1992). Intuitivamente hablando, cuando palabras como *"family"* o *"species"* aparecen en la entrada, la probabilidad para "ANIMAL" tiende a ser más grande que para "TOOLS"

Categoría	Palabras con características informativas
ANIMAL/INSECT (cat. 414)	Species (2.3), family (1.7), bird (2.6), fish (2.4), ...
TOOLS/MACHINERY (cat. 314)	Tool (3.1), machine (2.7), engine (2.6), blade (3.8), ...

Tabla 2.3. Categorías asociadas a características informativas

Sin embargo, en el trabajo de Justeson y Katz se selecciona la palabra w como indicador del sentido s siempre y cuando

[5] Estrictamente hablando, $P(w|s)$ y $P(w)$ deberían anotarse como $P(F_w = 1|s)$ y $P(F_w = 1)$, respectivamente, donde F_w es la característica que representa la existencia de la palabra w. Sin embargo, aquí se usa una notación simplificada.

w aparezca con más frecuencia con el sentido s que con otro sentido candidato. Particularmente, el objetivo de este trabajo es desambiguar el sentido de los adjetivos basándose en los antónimos. Por ejemplo, en la Tabla 2.4 se tienen nombres indicativos que pueden tener asociados distintos sentidos de un adjetivo. Hay que resaltar que en el trabajo de Justeson y Katz, el proceso de desambiguación en si mismo no depende de un modelo probabilístico, sino que los indicadores simplemente son usados como reglas de restricción.

Sentido	Nombres indicativos
Old ("No Nuevo")	car, motorbike, thing, computer, ...
Old ("No joven")	man, woman, cat, wine.....

Tabla 2.4. Nombres indicativos para un adjetivo

Otros investigadores (Pedersen et al., 1997; Ng y lee, 1996) utilizan, las propiedades morfológicas de las palabras polisémicas (singular/plural, el tiempo verbal), relaciones sintácticas asociadas con las palabras polisémicas, las características de las palabras y la colocación de estas en las distintas partes de la oración, como información relevante y útil para desambiguar el sentido de las palabras.

- A continuación, se calcula la probabilidad condicional $P(c|s_k)$ de cada rasgo o características para el sentido (o categoría) s_k. El modelo más simple que calcula esto, *Naive-Bayes method* combina hechos a partir de un gran número de características con una pobre eficiencia (Escudero et al., 2000c). Por lo tanto, si se desea aplicar sobre WSD, el contexto de la palabra w se describe en términos de v_j (palabras usadas como características contextuales para realizar la desambiguación) que aparecen en el contexto. Hay que resaltar que en el método de Naive-Bayes las características usadas se suponen condicionalmente independientes para cada uno de los sentidos dados, es decir $P(c|s_k)$ se calcula simplemente mediante el producto de $P(v_j|s_k)$ como se muestra en la fórmula 2.6.

$$P(c|s_k) = \prod_{v_j \ en \ c} P(v_j|s_k) \tag{2.6}$$

El método de Naive-Bayes explicado anteriormente, es el más simple para obtener la probabilidad condicional, y se aplicó a muchas investigaciones en WSD como son los casos de (Gale et al., 1992b; Mooney, 1996; Leacock et al., 1993; Pedersen y Bruce, 1997b; Escudero et al., 2000c; Suárez y Palomar, 2002). Naive Bayes se usa bastante para aprendizaje automático debido a su habilidad para combinar hechos o indicios a partir de un número de pruebas. Y es aplicable si el estado del mundo en el que basamos nuestra clasificación se describe como una serie de atributos.

El método de Naive-Bayes aplicado a WSD tiene dos consecuencias. La primera es que toda la estructura y ordenación lineal de las palabras en el contexto se ignora. Esto se conoce como el modelo "bolsa de palabras". Y la otra es que la presencia de una palabra en la bolsa es independiente de otra. Esto no es cierto. Por ejemplo, la palabra "presidente" es más común que aparezca en el contexto que contiene "elección" que en el contexto que contiene "poeta". Otro ejemplo ilustrativo, será el que a continuación se va a presentar en la Tabla 2.5 para los dos sentidos de *drugs*.

Sentido	Agrupación por sentidos
Medication	prices, patent, increase , prescription
Illegal substance	Abuse, alcohol, cocaine, traffickers

Tabla 2.5. Agrupación para "Drugs"

Traffickers es una buena palabra para el sentido de *illegal substance*, ya que la probabilidad condicional de que la palabra *traffickers* tenga el sentido *illegal substance* es superior a la presentada por *medication*. Por eso, si la palabra *traffickers* aparece en un contexto de *drug* posiblemente tendrá una probabilidad mas elevada con el sentido de *illegal substance* que con el de *medication*.

El trabajo de Pedersen (1997) presenta un modelo más complejo conocido como *"decomposable model"*, que considera diferentes características dependientes entre sí o recíprocamente. Sin embargo, este trabajo tiene la desventaja de que el número de parámetros a ser estimados es enorme, debido a que estos son proporcionales al número de combinaciones de los valores para las características que dependen entre sí. Además, esta técnica requiere grandes volúmenes de datos para estimar apropiadamente todos los parámetros. Para resolver este problema, Pedersen *et al.* (1997) proponen un método automático para identificar el modelo óptimo (alto rendimiento y poca estimación de parámetros), mediante la alteración iterativa del nivel de complejidad del modelo. Pero sus resultados demostraron que estos modelos generalmente no funcionan con el método de Naive-Bayes.

Otro trabajo de Escudero et al. (2000b) presenta la aplicación del algoritmo AdaBoost.MH a la tarea de desambiguar el sentido de las palabras. La idea principal de los algoritmos de boosting es combinar muchas hipótesis simples y seguras (llamadas clasificadores débiles) en un único clasificador para la tarea en cuestión. Los clasificadores débiles se entrenan secuencialmente, es decir, cada uno de ellos se entrena con los ejemplos más difíciles de clasificar por el clasificador débil que le precede. Los experimentos sobre un conjunto de 15 palabras polisémicas demuestran que la aproximación de boosting supera a la de *Naive-Bayes*.

2.2.3 Métodos basados en similitud semántica

Los humanos somos capaces de resolver problemas nuevos por medio de la semejanza con otros casos observados. Este proceso ha sido estudiado por varios trabajos (Bareiss, 1990; Aha et al., 1991; Weiss y Kulikowski, 1991; Kolodner, 1993), asignándole diferentes denominaciones. El cálculo de la similitud semántica entre un nuevo problema y los ejemplos que se tienen en los datos de entrenamiento ha sido un problema crítico. A continuación se expondrá una clasificación de estos métodos para

calcular la similitud semántica según se basen en la distribución estadística, obtenida a partir de los datos de entrenamiento, o en la dependencia de los recursos manuales como tesauros o diccionarios (recursos externos).

Métodos basados en similitud semántica a los datos de entrenamiento. Un método que pertenece a este apartado es el llamado los "k-vecinos más próximos" (en inglés, *"k-nearest neighbor" (k-NN)*), el cual procesa de la siguiente manera. Primero, se recuperan k ejemplos similares a la entrada a partir de los datos de entrenamiento. Después de eso, el sentido que tiene la mayor frecuencia en los k ejemplos se selecciona como la interpretación de la palabra de entrada. En el caso de que K=1, este método se denomina el "vecino más próximo" (en inglés, *"nearest neighbor"*). En este método, la palabra de entrada se desambigua simplemente con el sentido asociado con el ejemplo de mayor similitud. Ng (1997a) identificó automáticamente el valor óptimo de k sobre el rango de unos datos de entrenamiento dados, lo cual mejoró el rendimiento del método del "vecino más próximo".

Supongamos que cada ejemplo de entrenamiento así como la entrada (ejemplos) se representan por un vector de características definido por la fórmula 2.5, en la cual cada ejemplo se sitúa en un espacio N dimensional, donde la característica F_i se corresponde con el eje de coordenadas i. Una implementación de esto es el llamado *"vector space model"(VSM)*, el cual calcula la similitud entre dos ejemplos mediante el ángulo que forman los dos vectores obtenidos a partir de la representación de cada ejemplo. Schutze (1992) aplica VSM a la desambiguación del sentido de las palabras, mediante la representación de cada palabra con un vector que consta de la frecuencia de colocación de sus palabras. De esta forma, representamos a cada contexto con un vector, el cual es la suma de los vectores de las palabras que se relacionan con las palabras que aparecen en el contexto. Este método también necesita de algoritmos de agrupamiento automático para relacionar cada palabra polisémica con sus sentidos y representar vectores de sentidos. Finalmente, mediante la ecuación 2.7 se obtiene la similitud semántica entre la pa-

labra de entrada y cada agrupación del sentido de la palabra, seleccionando el sentido con máxima similitud.

$$Sim(x,y) = \frac{\overrightarrow{x} \cdot \overrightarrow{y}}{|\overrightarrow{x}| \cdot |\overrightarrow{y}|} \tag{2.7}$$

Leacock *et al.* (1993) compararon el método VSM, redes neuronales y el método de Naive-Bayes, y extrajeron la conclusión de que los dos primeros métodos superan muy ligeramente a este último en desambiguación.

Ng y Lee (1996) usaron en su método una medida diferente para calcular la similitud. Ellos utilizaron la distancia entre dos ejemplos sumando las distancias de los valores característicos asociados a esos ejemplos. Es decir, calcularon la similitud semántica a partir de la idea de que dos ejemplos son similares cuando tienen valores característicos que más o menos corresponden con la distribución estadística obtenida a partir de los datos de entrenamiento. La fórmula 2.8 muestra como se calcula la distancia de dos valores característicos f_1 y f_2 de una característica F.

$$dist(f_1, f_2) = \sum_i |P(s|F = f_1) - P(s|F = f_2)| \tag{2.8}$$

Donde $P(s|F = f_1)$ es la probabilidad condicional del sentido s para el valor f_1 de la característica F. Los resultados obtenidos en sus investigaciones indicaron que la selección de la característica es una tarea fundamental para los métodos basados en la similitud.

Otro investigador que obtiene la similitud entre dos ejemplos es Cho y Kim (1995), usando la entropía relativa (estima el grado mediante el cual dos distribuciones probabilísticas difieren) y aplicándola a la desambiguación del sentido de los verbos. Ellos usaron solo una característica, en la que cada ejemplo se representa mediante el nombre, que es objeto directo, asociado al verbo polisémico. La similitud entre ejemplos se calcula basándose en la distribución de los nombres asociados.

Métodos basados en similitud semántica mediante recursos externos. En realidad estos métodos son estadísticos, pero estiman la probabilidad de una clase en vez de una palabra

mediante WordNet. Los métodos clasificados en este subapartado se basan en la noción de qué palabras, próximas unas a otras en un tesauro, diccionario u otro recurso, tienen semejante significado. La mayoría de métodos (Li et al., 1995; Resnik, 1995a,b; Ribas, 1995b; Lin, 1997) predefinen heurísticamente la relación entre la similitud de nombres y la longitud del camino entre ellos, obtenida a partir de la estructura de los recursos externos.

2.2.4 Métodos basados en el uso de corpus bilingües

Estos métodos básicamente son estadísticos, pero sin embargo este apartado presenta otro parámetro de clasificación según la utilización de corpus bilingües. La principal característica de este tipo de métodos es que utilizan recursos bilingües para desambiguar los sentidos de las palabras, con el objetivo de probar ciertas técnicas de WSD sin la necesidad de etiquetar manualmente los sentidos en el texto. Estos métodos se basan en la observación de que diferentes sentidos de palabras en una lengua dada, pueden corresponderse con palabras distintas en otra lengua. Por ejemplo, la palabra en inglés *pen* es "bolígrafo" en español con el sentido de "instrumento para escribir" y "corral" con el sentido de "recinto para encerrar". Por lo tanto, si se dispone de un corpus alineado palabra a palabra, cuando se realiza una traducción de una palabra como *pen*, automáticamente se determina su sentido en español ("bolígrafo o corral").

En el trabajo realizado por Gale y Church (1991) se observó que las estructuras gramaticales y de párrafos en los dos corpus (francés e inglés) utilizados eran idénticas, ya que eran traducción directa el uno del otro. Por lo tanto aplicó una técnica automática de alineación de cada oración de un corpus con el otro, con el objetivo de obtener un corpus bilingüe alineado. Este trabajo se utilizó para descubrir cómo una palabra en una oración determinada se traducía al otro lenguaje, y por consiguiente la traducción de esa palabra a su sentido correspondiente. Por eso al aplicar esta técnica las palabras alineadas del corpus podían ser automáticamente etiquetadas con el sentido correcto. Las prue-

bas realizadas por Gale fueron sobre una única palabra (*bank*) y obtuvieron una precisión del 92%.

Otro autor que utilizó la correspondencia entre palabras de un diccionario bilingüe fue Dagan *et al.* (1994). Para explicar la idea de este método se utilizará un ejemplo. En inglés la palabra *interest* tiene dos traducciones en alemán: *Beteiligung* ("legal share") e *Interesse* ("attention", "concern"). Para desambiguar la palabra inglesa *interest*, hay que identificar la oración donde ésta aparece y buscar un corpus (alemán para nuestro ejemplo) con ejemplos de la oración. Si la oración aparece con una única traducción de la palabra *interest* en la segunda lengua, entonces se puede asignar el correspondiente sentido. Si la palabra *interest* aparece en la oración *showed interest*, la traducción en alemán de *showed* ("zeigen") aparecerá únicamente con *Interesse*. Por lo tanto, la palabra *interest* en la oración *showed interest* pertenece al sentido en alemán *"attention", "concern"*. Por otra parte, la única traducción frecuentemente utilizada de la oración *acquired interest* es *"erwarb eine Beteiligung"*, por lo tanto el uso de *interest* como el objeto de *acquire* corresponde al sentido, *legal share*.

Estos métodos de WSD basados en corpus bilingües tienen una limitación muy evidente, y es que solo pueden utilizarse en aplicaciones de traducción automática.

2.2.5 Métodos basados en propiedades discursivas

Los métodos basados en corpus pueden explotar dos propiedades muy poderosas del lenguaje humano como son: la propiedad de "Un sentido por discurso" y de "Un sentido por colocación" (Yarowsky, 1995). A continuación se explicarán detalladamente estas dos propiedades y su aplicación a los métodos de WSD.

- *"Un sentido por discurso"*. Esta propiedad se fundamenta en la idea de que el sentido de una palabra a ser desambiguada es altamente coherente con un documento dado. Con esta propiedad se está aplicando la observación de que en un determinado discurso, todas las ocurrencias de una misma palabra suelen

denotar siempre el mismo sentido. Por ejemplo, si estamos hablando de botánica, normalmente cuando aparezca la palabra "planta" en el discurso tendrá el sentido de "flora" en vez de "planta industrial".

- *"Un sentido por colocación"*. Esta propiedad se fundamenta en la idea de que si se agrupan palabras cercanas a la palabra a desambiguar, considerando la distancia, orden y relación sintáctica, éstas forman un grupo coherente con respecto a su sentido. Es decir, hay ciertos sentidos de ciertas palabras que quedan completamente determinados mediante una colocación, como por ejemplo, "materia gris". Al estar la palabra "gris" cerca de "materia", el sentido de esta queda claramente identificado como "tejido nervioso". Esta es la principal propiedad aplicada por la mayoría de los trabajos basados en procesos estadísticos para desambiguar el sentido de las palabras. Eso es debido a que estos dependen de que los sentidos de las palabras estén fuertemente relacionados con ciertas características contextuales, como por ejemplo, que haya otras palabras en la misma oración.

Estas dos características se aplicaron en el trabajo de Yarowsky (1995), el cual presentó un algoritmo no-supervisado que desambigua el sentido de las palabras en un corpus no etiquetado. El algoritmo evita la necesidad de utilizar datos de entrenamiento etiquetados manualmente, ya que utiliza las dos propiedades comentadas anteriormente. El algoritmo consta de un procedimiento basado en corpus que reúne características del contexto local, para más tarde poder utilizarse en WSD. Yarowsky demostró que este algoritmo es muy eficaz, ya que para diferentes versiones obtuvo una precisión que oscilaba entre 90,6% y 96,5%. Cuando se aplicaba la característica del discurso el ratio de error se reducía en un 27%.

2.3 Métodos basados en Conocimiento

El trabajo que se presenta en esta Tesis está clasificado como un método basado en el conocimiento. Por tal motivo, esta Tesis se

basa en algunos conceptos y técnicas que se han utilizado en los trabajos clasificados en este apartado.

Los métodos de desambiguación supervisados requieren de grandes conjuntos de datos de entrenamiento. Normalmente, los algoritmos de desambiguación basados en corpus usan datos de entrenamiento etiquetados manualmente. Es decir, cada palabra ambigua perteneciente al conjunto de datos de entrenamiento tendrá anotada una etiqueta semántica con su sentido apropiado. Esto supone una barrera bastante importante para desambiguar el sentido de las palabras por la enorme dificultad y el elevado coste para etiquetar manualmente las grandes cantidades de información requeridas para desambiguar el sentido de las palabras. Por ejemplo, (Ng, 1997b) estimó que un hombre con dedicación exclusiva tardaría 16 años para construir un corpus etiquetado semánticamente para el inglés. Otro factor a tener en cuenta es la diversidad de lenguas, por lo que se debería realizar la anotación de los corpus para cada una de las lenguas. Debido a los factores comentados anteriormente (cobertura, anotación para cada lengua) los trabajos más recientes en WSD se han enfocado para reducir la necesidad de supervisar los métodos basados en corpus, así como el coste de adquisición de los corpus etiquetados semánticamente. Debemos nombrar los trabajos de Mihalcea y Moldovan (1998) y de Agirre y Martinez (2001a) para encontrar ejemplos de sentidos de WordNet en la Web de forma automática.

Durante los años ochenta, los recursos léxicos como diccionarios, tesauros y léxicos basados en conocimiento se pusieron a disponibilidad de los investigadores facilitando la obtención de conocimiento a partir de los mismos sin necesidad del etiquetado manual. Estos recursos proporcionan una fuente de información muy importante en los métodos de WSD basados en conocimiento, o también conocidos en inglés como *knowledge-driven WSD.*

Los primeros métodos de desambiguación basados en conocimiento, estuvieron muy influenciados por las técnicas que se utilizaban en la Inteligencia Artificial (IA). Por tal motivo, los métodos de desambiguación basados en conocimiento se han clasificado en base a dos criterios: *Métodos basados en técnicas de IA* y *Métodos basados en recursos externos..*

A continuación se presentan las investigaciones de mayor relevancia en la WSD basada en conocimiento según los criterios de clasificación expuestos anteriormente.

2.3.1 Métodos basados en técnicas de IA.

Los primeros métodos basados en IA empezaron abordando el problema de la comprensión del lenguaje. Por eso, los trabajos en WSD siguiendo un enfoque de IA se realizaron en el contexto de sistemas grandes, con la intención de obtener una comprensión completa del lenguaje. Para realizar esta tarea se tenía que obtener conocimiento sintáctico y semántico, lo cual posteriormente fue explotado por la tarea de WSD. A continuación, se van a explicar los métodos más representativos o característicos de la WSD basados en técnicas de IA.

Las redes semánticas como método de desambiguación. Las redes semánticas[6] son representaciones gráficas de conocimiento compuestas de nodos y conexiones (arcos o flechas) entre esos nodos. Los nodos representan objetos o conceptos y las conexiones representan las relaciones entre esos nodos.

Un ejemplo de red semántica sería el mostrado en la figura 2.3.

Las redes semánticas permiten representar el sentido de las palabras. En el trabajo de Masterman (1962) se utilizó una red semántica para derivar la representación de oraciones en una interlingua (entre varias lenguas) formada de conceptos esenciales del lenguaje y se aplicó al campo de la traducción automática. La distinción de sentidos se hacía implícitamente, mediante la elección de representaciones que sean grupos de nodos fuertemente relacionados en la red.

En el trabajo de Quillian (1968; 1969) se presenta la construcción de una red semántica, que incluía enlaces entre palabras (tokens) y conceptos (tipos). A dichos enlaces se les etiqueta con varias relaciones semánticas o mediante asociaciones entre palabras. La red es creada a partir de las definiciones de diccionario, pero se mejora mediante conocimiento etiquetado manualmente.

[6] Las redes semánticas provienen de los primeros trabajos sobre grafos para la representación del conocimiento (Roberts, 1973; Selz, 1922).

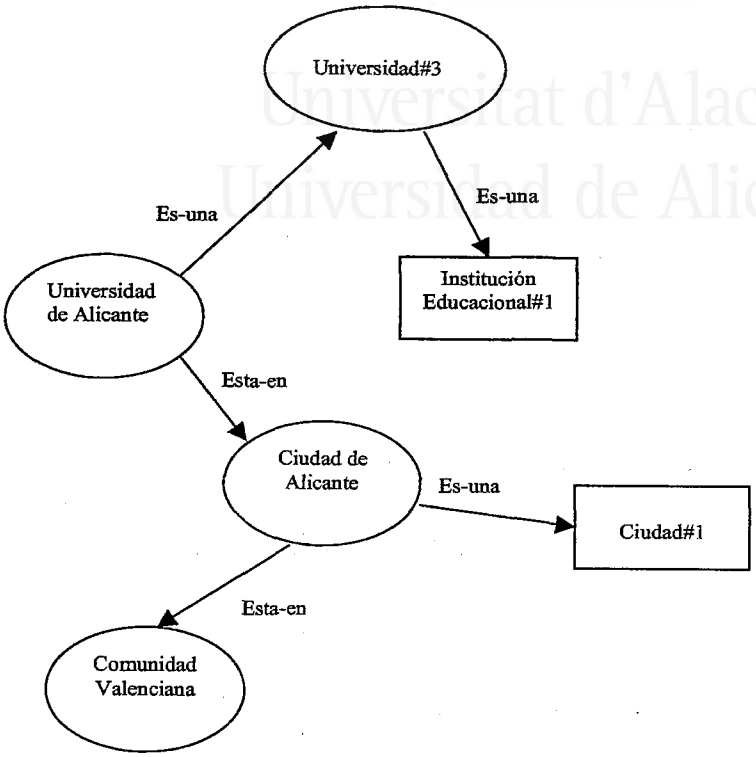

Figura 2.3. Estructura de una Red Semántica

Cuando dos palabras se presentan a la red, el programa desarrollado por Quillian simula la activación gradual de nodos (conceptos) a lo largo de un camino de enlaces, originado a partir de cada palabra de entrada. En este trabajo la desambiguación se consigue cuando sólo un nodo de un concepto asociado con la palabra de entrada dada se involucra en el camino más directo encontrado entre dos palabras de entrada.

Por otro lado, estos métodos también se basaron en el uso de "marcos de reglas" (*frames*). Los "marcos de reglas" son estructuras de datos que contienen información de palabras y sus relaciones. Hayes (1977a; 1977b; 1978) utilizó la combinación de una red semántica y de "marcos de reglas", de tal manera que construyó

una red semántica formada por nodos, que representan los sentidos de los nombres, y por enlaces que representan los sentidos de los verbos. Sin embargo, para los "marcos de reglas" se aplicaron las relaciones IS-A y PART-OF a la red. El trabajo que desarrolló Hayes demuestra que los homónimos pueden ser desambiguados con bastante éxito usando esta aproximación pero, sin embargo, para otros tipos de polisemia no se alcanza la misma precisión.

En el trabajo de Hirst (Hirst, 1987) también se usaron redes de "marcos de reglas" y se introdujo el mecanismo denominado *polaroid words*. Este mecanismo consiste en eliminar progresivamente los sentidos no adecuados, basándose en los indicios sintácticos (suministrados por el analizador) y en las relaciones semánticas encontradas en la red de "marcos de reglas".

Debemos mencionar que todos estos trabajos no eran a gran escala, sino todo lo contrario. Es decir, se trabajaba sobre "ejemplos juguete".

Modelos conexionistas como métodos de desambiguación. La idea de los modelos conexionistas aplicados a la desambiguación del sentido de las palabras consiste en un proceso en el cual la introducción de un cierto concepto facilitará el procesamiento de subsiguientes conceptos introducidos que están relacionados semánticamente. Esta idea se aplica en los "modelos de activación extendida" de Anderson (1983) donde los conceptos de una red semántica se activan por su uso y ésta se propaga por los nodos de la red.

Cottrell y Small (1983) propusieron e implementaron un modelo de activación extendida para desambiguar el sentido de las palabras en el cual cada nodo de la red representa una palabra específica o concepto. Otros investigadores como Waltz y Pollack (1985) anotaron manualmente conjuntos de características semánticas correspondientes a duración de eventos (minuto, hora, día, etc), localización (ciudad, país, etc), restricciones semánticas (animado/inanimado, etc) y otras más, en sus redes. Posteriormente, estos conjuntos de características, definidas manualmente, se utilizaron para activar un contexto determinado que servía para desambiguar el sentido de una palabra de entrada. Otro trabajo es el de Bookman (1987), el cual describe un proceso dinámico

en el que las características son automáticamente activadas por el texto precedente. Todos los trabajos propuestos hasta ahora utilizan modelos en donde un nodo corresponde a un único concepto, denominándose modelos locales. Sin embargo, Kawamoto (1988) propuso modelos denominados distribuidos, los cuales requieren una fase de aprendizaje que usa ejemplos previamente desambiguados.

Por otra parte, la gran dificultad para anotar manualmente conjuntos de características semánticas hace que solamente se apliquen estos modelos a una parte muy reducida del lenguaje. Y por lo tanto, los procedimientos de desambiguación que forman parte de estos modelos alcanzan su efectividad cuando se aplican a "ejemplos juguete", es decir a conjuntos de datos pequeños con un contexto muy limitado. Finalmente, todo lo comentado anteriormente desemboca en modelos que no tienen una efectividad elevada sobre textos reales.

Las preferencias semánticas como método de desambiguación. Las preferencias semánticas son representaciones de conocimiento en donde las expectativas semánticas se almacenan como restricciones manejables o preferencias. Se usa el término preferencias en contraposición a reglas, porque las preferencias se pueden anular en metáforas y otros significados extendidos. Por ejemplo, la palabra "comer" toma argumentos que no se corresponden con alimento en la oración "el cantante se comió unas palabras". Así, el uso de las preferencias semánticas permite a los sistemas de conocimiento combatir estos casos, ya que aportan flexibilidad para manejar la entrada de algo inesperado. Es decir, permite tomar la acción de asumir automáticamente que cualquier cosa que se sale de la competencia semántica del sistema, no tiene sentido. Otro ejemplo, es la palabra ambigua "puente" que puede tomar el sentido de "estructura" o de "aparato dental". Si se tiene la oración, "yo pinté el puente", parece claro que el sentido de esta palabra va dirigido hacia "estructura". Pero, imaginemos que se tiene una restricción de selección para "pintar" como que el TEMA="objeto físico". Por lo tanto, ambos sentidos de "puente" pertenecen a la categoría de "objeto físico" y por lo tanto satisfacen la restricción. Por lo tanto, no se seleccionaría el sentido

correcto de las palabras aplicando esa restricción. Este problema se puede resolver mediante la utilización de técnicas que utilicen preferencias semánticas.

Las investigaciones de (Wilks, 1968, 1969, 1973, 1975d,c,b,a) son unos de los primeros trabajos que están específicamente destinados para tratar el problema de la desambiguación del sentido usando las preferencias semánticas. Así, el trabajo de Wilks (1975c) describe que las preferencias semánticas especifican de forma precisa a las restricciones de selección. Con el objetivo de realizar combinaciones de términos en una oración para que se pueden suavizar sus restricciones cuando no se dan las restricciones preferidas de una palabra. Por consiguiente se habilita el manejo de la metáfora. Por ejemplo, en la oración "Mi coche bebe gasolina" las restricciones de "beber" prefieren un sujeto animado pero permiten a uno inanimado. Sin embargo, el trabajo de Boguraev (1979) considera que hay que tener en cuenta los sentidos de los verbos si se quieren utilizar la preferencias semánticas para WSD. Para ello usó una combinación de restricciones seleccionales, preferencias, frames, etc, con objeto de mejorar el método propuesto por Wilks. Otros autores más recientes como (Resnik, 1993a; Ribas, 1995b; Li y Abe, 1995) también tratan el problema de la desambiguación del sentido aplicando restricciones seleccionales.

Las redes de discriminación como método de desambiguación. Una red de discriminación permite clasificar un objeto en una estructura según sus propiedades. Las redes de discriminación se pueden usar para sugerir una sucesión de tests que aplicados a un objeto nos permiten clasificarlo dentro de una estructura.

A continuación se va a presentar un ejemplo de cómo las redes de discriminación permiten organizar la información de ciertas tareas. Consideramos un juego con dos jugadores en el que el primer jugador (A) piensa en un animal, y el segundo jugador (B) tiene que adivinar cual es. El jugador (A) solo puede responder a las preguntas del jugador (B) con respuestas de sí-no. Una simple muestra del juego sería el mostrado en el ejemplo 2.

(2) **B:** Puedes comerlo?

A: No
B: Tiene piernas?
A: Si
B: Es un caballo?
A: No
B: Tiene cuatro piernas?
A: No
B: Es un hombre?
A: Si

A continuación, se expone el ejemplo 2.4, el cual presenta una red de discriminación. Como se puede ver en este ejemplo, el constituyente básico de esta red de discriminación es la representación de una pregunta y el texto adicional que puede ser relevante de acuerdo a si la respuesta es afirmativa o negativa. El otro tipo de constituyente en esta red es para representar los hechos como *Si tiene 4 piernas y Si puedes montarlo entonces el objeto debe ser un caballo.*

En este ejemplo hay dos corchetes después de [objeto una vaca]. El primero cierra la respuesta negativa para la pregunta [Puedes montarlo] y el segundo cierra la respuesta afirmativa para la pregunta [Tiene cuatro piernas].

Este enfoque lo que pretende es realizar una discriminación de los sentidos como en (Small y Rieger, 1982; Small, 1983; Small et al., 1988). Okumura y Tanaka (1990) proponen una buena estrategia para la desambiguación semántica mediante las redes de discriminación. Esta consiste en aumentar las restricciones obtenidas durante el proceso de análisis de una oración, y desambiguar lo más pronto posible el sentido mediante esas restricciones. En este trabajo se presenta el modelo computacional para el análisis del lenguaje denominado *Incremental disambiguation model.*

Uso de información de diversa naturaleza como método de desambiguación. Los métodos de desambiguación que se presentan aquí se basan en usar una gran variedad de información de diversa naturaleza como análisis sintáctico, restricciones de selección, etc.

```
[cuestión
   Tiene cuatro piernas
   [cuestión
      [Puedes montarlo]
      [objeto un caballo]
      [objeto una vaca]]
   [objeto un hombre]]

Si tiene 4 piernas entonces
    Si puedes montarlo entonces
        Es un caballo
    Sino
        Es una vaca
Sino
    Es un hombre
```

Figura 2.4. Red de discriminación

Dahlgren (1988) presenta un sistema de comprensión del lenguaje natural que incluye un método para desambiguar el sentido de las palabras, el cual utiliza información de diversa naturaleza como información sintáctica, expresiones determinadas y razonamiento utilizando el sentido común. Este último tipo de información, es decir el razonamiento, solamente lo aplica cuando fallan los otros dos métodos en la obtención del resultado. De este trabajo se obtiene la conclusión de que el razonamiento normalmente necesita una ontología para encontrar antecesores comunes a las palabras que forman parte del contexto. El trabajo anterior se anticipa a los resultados obtenidos en (Resnik, 1993b,c, 1995a), donde determinan que la semejanza ontológica trae consigo un antecesor común en la ontología. Con esto se prueba que esta idea es muy importante para desambiguar además de demostrar que las restricciones de selección del verbo son un recurso muy importante para realizar la desambiguación del sentido de los nombres.

El trabajo de Agirre y Martinez (2001b) presenta una extensión de los modelos estadísticos previos a preferencias de selección clase-a-clase. También, presenta un modelo que aprende preferencias de selección para clases de verbos. Estas preferencias

extraídas se evaluaron en WSD con un conjunto de palabras y documentos del SemCor.

2.3.2 Métodos basados en recursos externos estructurados.

El fundamento de los métodos basados en recursos externos consiste en intentar usar el contexto de la palabra a ser desambiguada junto con la información, que se posea en recursos léxicos externos, de cada uno de los sentidos de las palabras con el objetivo de asignar el sentido correcto. Esta información para cada uno de los sentidos se obtiene a partir de recursos de conocimiento externos (diccionarios, tesauros, bases de conocimiento léxicas). La clasificación que se propone para este tipo de métodos se basa en el tipo de recurso externo estructurado utilizado. Así la subclasificación para este tipo de métodos es la siguiente: *diccionarios de uso convencional (MRD´s)*, *tesauros*, y *bases de conocimiento léxicas.*

A continuación se detallan las características de cada uno de los métodos expuestos anteriormente.

Métodos basados en conocimiento extraído de diccionarios externos. Un diccionario electrónico (Machine Read Dictionary (MRD)) es un diccionario convencional en soporte electrónico, que detalla de forma ordenada el vocabulario de una lengua para conocer sus significados.

Aunque los diccionarios contienen multiples inconsistencias (Atkins y Levin, 1988; Kilgarriff, 1994; Rigau et al., 1998) y no se crearon para su explotación en ordenadores, lo cierto es que proporcionan una fuente de información muy valiosa para la distinción de los sentidos de las palabras, a través de sus definiciones. Por este motivo se utilizan como recurso externo en bastantes métodos de WSD.

El método que ha servido de base o de punto de partida para todos los restantes métodos de WSD usando MRD´s ha sido el de Lesk (1986). Este asociaba a cada sentido del diccionario una lista de palabras que aparecían en la definición de ese sentido. Entonces

la desambiguación se realizaba seleccionando el sentido de la palabra cuya lista contenía el mayor número de coincidencias con las claves de las palabras de su contexto. El inconveniente que tienen este método es que es muy sensible a la redacción exacta de cada definición, ya que la ausencia o presencia de una palabra puede alterar radicalmente los resultados, por lo que dependen en gran medida de las palabras elegidas para describir las definiciones.

El trabajo de Wilks *et al.* (1990; 1996) intentó mejorar el conocimiento asociado con cada sentido, mediante el cálculo de la frecuencia de aparición de las palabras en la definición de los textos del diccionario *Longman´s Dictionary of Contemporary English* (LDOCE). Esta métrica se apoya con un vector conocido en inglés como *"Co-ocurrence Vector"*, el cual relaciona cada palabra con su contexto. Como las definiciones en el diccionario LDOCE han sido escritas utilizando un reducido vocabulario (2781 palabras) entonces las frecuencias se limitan a esas palabras. A este vector, 2.9, se asociará un valor a cada palabra del reducido vocabulario encontrado en LDOCE (N en la ecuación 2.9 es igual al tamaño del vocabulario reducido (2781)), el cual representará la frecuencia de aparición (ver fórmula 2.10) para la palabra w y la palabra i-th en el mismo vocabulario. Con la ecuación 2.10 se obtienen cada uno de los valores del vector de ocurrencias, los cuales son la frecuencia de las apariciones u ocurrencias.

$$\overrightarrow{v_w} = (v_0^w, \cdots, v_N^w) \tag{2.9}$$

$$v_i^W = f_{w,z_i} \tag{2.10}$$

Igual que se calcula la relación entre dos palabras, se puede calcular matemáticamente la relación entre los dos vectores correspondientes, mediante por ejemplo el coseno (ver fórmula 2.12). Finalmente, un sentido puede caracterizarse agregando todos los vectores correspondientes a las palabras en su definición (Wilks et al., 1990) (aplicando la fórmula 2.11). De esta forma se obtiene un vector para cada sentido del diccionario.

$$\overrightarrow{v_a} = \sum_{n \in def(a)} \overrightarrow{v_w} \tag{2.11}$$

Se pueden aplicar distintas medidas de similitud entre vectores en un espacio multidimensional donde (las dimensiones las forman las palabras) como por ejemplo la fórmula del coseno 2.12, que mide la relación entre dos vectores.

$$sim(v,w) = \cos(\overrightarrow{v},\overrightarrow{w}) = \frac{\sum_{k=1}^{N}(v_k w_k)}{\sqrt{\sum_{k=1}^{N} v_k^2 \sum_{k=1}^{N} w_k^2}} \qquad (2.12)$$

La evaluación del método se hizo sobre una única palabra "bank" obteniéndose una exactitud del 45% en la identificación del sentido y del 90% en la identificación de palabras homógrafas.

Varios autores como (Krovetz y Croft, 1989; Guthrie et al., 1991; Slator, 1992; Cowie et al., 1992; Liddy y Paik, 1993) intentaron mejorar los resultados obtenidos en investigaciones anteriores con el uso de campos suplementarios de información en la versión electrónica del LDOCE, en particular, los campos *box code* y *subject code* para cada sentido. El campo *box code* incluye primitivas tales como "Abstracto", "Animado", "Humano", etc., y codifican tipos de restricciones de nombres, adjetivos y sobre los argumentos de los verbos. El campo *subject code* codifica un conjunto de primitivas que clasifican muchos de los sentidos de las palabras por temas ("Ingeniería", "Botánica", "Economía", etc).

El método de desambiguación empleado por Guthrie *et al.* (1991) usando los *subject codes* de LDOCE es similar al de Wilks, con la salvedad que, en el proceso de expansión de las definiciones, una definición que está asignada a una categoría sólo puede expandirse con palabras concurrentes presentes en otras definiciones asignadas a la misma categoría. Los resultados de la evaluación no se disponen, pero Cowie *et al.* (1992) conjuntamente con Guthrie mejoraron el método obteniendo unos resultados del 47 % en la distinción de sentidos y del 72 % para las palabras homógrafas.

Otro trabajo es el de Richardson (1997), el cual presentó el desarrollo de una red semántica a partir de las definiciones de dos diccionarios (LDOCE y Webster´s 7th Gove) mediante el análisis sintáctico de sus definiciones con el objetivo de extraer relaciones semánticas y asignarles un peso basado en la frecuencia. Como las palabras no tienen asignados sus sentidos en las definiciones, en la red semántica es imposible la relación entre dos sentidos. Por lo

tanto, en vez de eso se usaron caminos de enlace para implementar las relaciones entre palabras. En definitiva, la idea consiste en que dos palabras estarán muy relacionadas si hay muchos caminos de enlace entre ellas. Pero esos caminos de enlace no tienen la misma importancia por lo que hay que medir la importancia de cada uno de ellos con respecto a cada tipo de relación.

Métodos basados en conocimiento extraído de tesauros. Los tesauros proporcionan información sobre las relaciones entre palabras del inglés (*Roget´s International Thesaurus*). La relación que se utiliza más comúnmente es la de sinonimia.

El tesauro Roget se suministró en formato electrónico (disponible via ftp anónimo desde varios lugares[7] y, a partir de la década de los 50 se ha utilizado en una amplia variedad de aplicaciones, tales como recuperación de información, traducción automática y análisis de contenidos. Un tesauro está formado por un conjunto de 1000 categorías y para cada una de ellas un grupo de palabras relacionadas entre sí. Por lo tanto, cada ocurrencia de una misma palabra en diferentes categorías del tesauro, representa diferentes sentidos o significados de la palabra. Así, este recurso se puede utilizar para desambiguar el sentido de las palabras, ya que palabras pertenecientes a una misma categoría están semánticamente relacionadas (Yarowsky, 1992) y dichas categorías se corresponde con los sentidos de esas palabras.

Uno de los primeros algoritmos basados en tesauros para desambiguar una palabra fue propuesto por Walker (1987), y este utilizó la idea de que cada sentido y cada palabra se asignan a una o más categorías o temas en el diccionario. Por lo tanto para desambiguar una palabra, se extraían del diccionario sus categorías, para posteriormente calcular la frecuencia de aparición de las palabras de la frase en las categorías y finalmente seleccionar el sentido para aquella categoría más frecuente.

Un problema con el algoritmo es que una categorización general de palabras en temas es inapropiado para un dominio particular. Por ejemplo, la palabra "ratón" en un tesauro puede aparecer como mamífero o como dispositivo electrónico. Sin embargo, en

[7] `ftp.clr.nmsu.edu:/CLR/lexica/roget-1911`

un manual de ordenadores rara vez aparecerá la palabra "ratón" como la categoría mamífero. Otro inconveniente encontrado en este algoritmo es la cobertura en una categorización general de temas. Por ejemplo, si se busca el nombre propio *Luís Figo* en un thesaurus de los años 60 no se encontraría, sin embargo la ocurrencia de este nombre nos daría un indicio para clasificarlo como "deporte". Para resolver estos problemas, el trabajo de Yarowsky (1992) añade las palabras más frecuentes de un corpus a las categorías según el contexto, por ejemplo, la ocurrencia *Luís Figo* de un corpus aparecería más frecuentemente en un contexto de deportes, por lo tanto se añadiría a la categoría deporte.

Yarowsky (1992) desarrolló un algoritmo para desambiguar el sentido de las palabras, usando el tesauro Roget y la Enciclopedia Multimedia Grolier. Este algoritmo toma como sentidos las categorías semánticas dadas en el tesauro Roget, en donde cada categoría semántica se compone de un conjunto de palabras relacionadas. Utilizó una amplia variedad de categorías para cubrir distintas áreas, como por ejemplo *herramientas/maquinaria* o *animales/insectos*. Y para saber cual es el contexto típico de cada categoría, reunía contextos para cada palabra en una categoría de la enciclopedia Grolier. Finalmente, a partir de todas las palabras en el contexto de la categoría se elegían las más significativas[8] mediante una medida estadística llamada *saliency* y que se muestra en la fórmula 2.13. Esta fórmula calcula la probabilidad de una palabra que aparece en el contexto de una categoría Roget dividido por su probabilidad general en el corpus.

$$saliency(w) = \log \frac{P(w|c)}{P(w)} \tag{2.13}$$

En este trabajo la relación de palabras no se define explícitamente, sino que se usa implícitamente como un método para etiquetar palabras con las categorías de Roget´s. Sin embargo, es posible inferir la relación entre palabras o sentidos a partir de la medida dada.

[8] Se entiende por significativas a aquellas palabras que aparecen más a menudo con el contexto de una categoría

Los experimentos realizados demuestran que el algoritmo tiene una alta precisión cuando las categorías de los tesauros y los sentidos se corresponden con los temas, por ejemplo la palabra *bass* tiene los sentidos *music y fish* y las categorías Roget son *music y animal.*

En definitiva, la idea básica en la desambiguación basada en thesaurus es que las categorías semánticas de las palabras en un contexto, determinan la categoría semántica del contexto en su totalidad, y por lo tanto determina el sentido de las palabras (Manning y Schütze, 1999).

Métodos basados en conocimiento extraído de bases de conocimiento léxicas. A mediados de los años 80, los investigadores empiezan a construir manualmente bases de conocimiento a escala real, como por ejemplo WordNet (Miller et al., 1990; Fellbaum, 1998), CyC (Lenat, 1995), ACQUILEX (Briscoe, 1991; Verdejo, 1994) y COMLEX (Grishman et al., 1994). Sin embargo, WordNet es la base de conocimiento en inglés más utilizada para desambiguar el sentido de las palabras, por varias razones que se detallarán a continuación:

1. WordNet combina las características de otros recursos explotados comúnmente en el trabajo de desambiguación: ya que incluye definiciones de términos para sentidos individuales como en un diccionario; define *Synsets*, como conjunto de palabras sinónimas que representan un único concepto léxico, y los organiza en una jerarquía conceptual como un tesauro. También, incluye otros tipos de relaciones léxicas y semánticas (hiperonimia, hiponimia, meronimia, etc), que proporcionan el conjunto más amplio de información léxica en un único recurso.
2. WordNet se diseñó para ser usado por programas, por lo tanto, no tiene muchos de los problemas asociados de los diccionarios electrónicos (MRD´s) comentados anteriormente. Y por último y quizás la razón más convincente para el uso de WordNet es que es el primer recurso con una amplia cobertura que se distribuye gratis y tiene una amplia disponibilidad.

La característica fundamental de estos métodos es que la asociación de palabras a los sentidos se cumple dependiendo de un recurso de conocimiento externo (por ejemplo, WordNet). Para WSD existen diferentes métodos de trabajo que utilizan este enfoque como puede verse en el trabajo de Ide y Veronís (Ide y Véronis, 1998).

Sin embargo, como el trabajo que se presenta en esta Tesis está clasificado más concretamente como un método basado en el conocimiento y que utiliza una base de conocimiento léxica, en particular WordNet. Las investigaciones más importantes clasificadas con este enfoque se presentarán con mayor detalle en el siguiente apartado, ya que el método propuesto en esta Tesis se basa en algunos conceptos y técnicas previamente utilizados en estos trabajos.

2.3.3 Métodos basados en conocimiento extraído de bases de conocimiento léxicas

El método tradicional para medir la similitud en una red semántica consiste en medir la longitud del camino entre dos conceptos (Rada et al., 1989). La fórmula 2.14 de distancia entre dos conceptos A y B, que proponen, se define como la longitud del camino más corto de la relación IS-A que unen a ambos conceptos. La medida de la distancia debería ser más pequeña para dos conceptos fuertemente relacionados, y viceversa.

$$dist(A,B) = \min_{p \in camino(A,B)} longitud(p) \tag{2.14}$$

La distancia conceptual, que propusieron Agirre *et al.* (1994), entre dos conceptos (a y b en la fórmula 2.15) se obtiene por el camino (p) más corto. Para realizar el cálculo de la longitud se hace de una forma especial: para cada concepto C_i en el camino se añadirá la inversa de su profundidad en la jerarquía, donde $a = C_0$ y $b = C_n$.

$$dist(a,b) = \min_{p \in camino(a,b)} \sum_{C_i \in p} \frac{1}{profundidad(C_i)} \tag{2.15}$$

Uno de los primeros autores que explotó la información léxica existente en WordNet para WSD aplicado a la recuperación de información fue Voorhees (1993). Para ello utilizó las relaciones de hiponimia para nombres en WordNet. Y definió la "capucha" de un sentido (en inglés, *hood*) o *synset* como la categoría de sentidos asociado a él. Algo similar a la representación de categorías realizada sobre el Roget en los métodos que ya hemos comentado anteriormente, pero adaptado al *synset*. Para definir la "capucha" de un *synset* se considera el conjunto de *synsets* y los enlaces de hiponimia de WordNet como el conjunto de vértices y arcos dirigidos de un grafo. Voorhees propone un criterio basado en la frecuencia de aparición de las "capuchas" en la colección de documentos y en el texto considerado. Así, dada una palabra en un documento y sus "capuchas" asociadas (una para cada *synset*), se le da mayor peso a las "capuchas" que son más referenciadas en el texto respecto a la colección. Esto permite asignar un mayor peso a las "capuchas" que son más referenciadas en un documento, y que deben representar mejor el significado concreto del término en el documento. Cuando se tienen representados los documentos por sus "capuchas" con sus pesos, se define una función de similitud como la del modelo del espacio vectorial. El concepto de "capucha" es un intento de distinguir los sentidos de manera menos fina que los synsets de WordNet y de manera más fina que las 10 jerarquías de nombres de WordNet. Sus resultados indican que su técnica no es un método fiable para distinguir los *synsets* de WordNet.

Otro autor, como Sussna (1993) utilizó la red semántica de nombres de WordNet como fuente de información para su técnica de WSD. La red permitía realizar el cálculo de la distancia semántica entre dos palabras cualesquiera pertenecientes a dicha red, por lo tanto también se podía calcular la distancia semántica para cada nombre de un conjunto de textos de entrada con el objetivo de desambiguarlos. Para realizar esto, Sussna asignó pesos según los distintos tipos de relación (sinonimia, hiperonimia, etc) de la red semántica, reflejando la similitud semántica expresada por esa relación. La distancia semántica entre dos nodos (sentido de la palabra) se calculaba sumando los pesos de las relaciones

del camino más corto entre los dos nodos y se seleccionaban los sentidos que minimizaban la distancia. Resumiendo, la hipótesis de este trabajo consiste en que para un conjunto dado de nombres cercanos en un texto, se eligen los sentidos que minimicen la distancia entre ellos. Hay que resaltar que los resultados de la WSD obtenidos por Sussna son significativamente mejores que el anterior autor. Este trabajo es interesante porque es uno de los pocos que utiliza otras relaciones léxicas de WordNet además de la hiponimia.

Agirre y Rigau (1995; 1996) presentaron un método para resolver la ambigüedad léxica de nombres, que se basa en la noción de densidad conceptual (extensión de la distancia conceptual empleada por (Rada et al., 1989; Sussna, 1993)) entre conceptos[9]. El sistema necesita saber cómo se agrupan las palabras en clases semánticas y cómo se organizan jerárquicamente, y para este propósito usaron la taxonomía semántica de nombres de WordNet. El funcionamiento de este método para WSD es el siguiente:

Se toma una ventana de contexto[10] W, que se irá moviendo a lo largo de todo el documento nombre a nombre, con el objetivo de desambiguar en cada paso el nombre que está en el centro de la ventana y considerar como contexto al resto de nombres dentro de la ventana. Posteriormente, el proceso que desambigua una palabra dada w del contexto C, siendo el contexto el resto de palabras de la ventana, utiliza 5 pasos que proceden de la siguiente manera:

Paso 1: Se presentan los nombres presentes en la ventana, sus sentidos y sus hiperónimos.

Paso 2: Se calcula la densidad conceptual de cada concepto en WordNet, de acuerdo a los sentidos contenidos en su subjerarquía.

Paso 3: Se selecciona el concepto c con mayor densidad conceptual.

Paso 4: Si w tiene un único sentido por debajo del concepto c, esta ya ha sido desambiguada, pero si no la tiene, entonces to-

[9] Un concepto es análogo a un *synset* en WordNet.

[10] Se entiende por ventana de contexto al grupo de palabras (10 palabras, 20, etc) que forman parte del texto a desambiguar.

davía está pendiente de desambiguar. Pero se podrán eliminar todos los demás sentidos de la palabra w que no estén bajo la jerarquía de c.

Se procede entonces a calcular la densidad para el resto de sentidos de las palabras de la ventana, y se continua desambiguando las palabras que hay a la izquierda en el contexto C (se vuelve al paso 2, 3, 4). Cuando no es posible realizar una desambiguación mejor, se procesan los sentidos a la izquierda de la palabra w y se muestra el resultado. Hay que aclarar que en algunos casos el método de desambiguación devuelve varios sentidos posibles para una misma palabra.

Hasta ahora se ha explicado el método utilizado para WSD pero no se ha presentado nada de lo que aporta la densidad conceptual y como se realiza el cálculo de la misma. A continuación se presentarán estas dos actividades detalladamente.

La densidad conceptual aporta una medida para calcular la cercanía de los sentidos de dos palabras en una red jerárquica estructurada. Así, lo que se pretende obtener con la medida de distancia conceptual entre conceptos de la jerarquía es:

1. La longitud del camino más corto que conecte los conceptos.
2. La profundidad en la jerarquía. Conceptos profundos en la jerarquía deberían considerarse más cercanos que los de partes altas.
3. La densidad de conceptos en la jerarquía. Áreas con muchos conceptos deberían considerarse más próximas que áreas con menos conceptos.
4. La medida debería ser independiente del número de conceptos que se miden.

La fórmula de densidad conceptual de Agirre y Rigau (1995) pretende ampliar los cuatro criterios vistos anteriormente y su cálculo es el siguiente.

Dado un concepto c (en lo alto de la jerarquía), y $nhyp$ y h (número de hiperónimos por nodo y altura de la subjerarquía, respectivamente), la densidad conceptual para c cuando su subjerarquía contiene un número m de sentidos de las palabras a desambiguar se calcula con la fórmula 2.16:

$$CD(c,m) = \frac{\sum_{i=0}^{m-1} nhyp^{i^{0.20}}}{descendientes_c} \quad (2.16)$$

donde el denominador viene dado por la fórmula 2.17:

$$descendientes_c = \sum_{i=0}^{h-1} nhyp^i \quad (2.17)$$

El numerador expresa el área esperada para una subjerarquía conteniendo m sentidos de las palabras a ser desambiguadas y el divisor es el área estimada. Es decir, proporciona la razón entre los sentidos ponderados de los sentidos a ser desambiguados bajo c y el número de sentidos descendientes estimados del concepto c. En definitiva, la fórmula 2.16 captura la relación entre los sentidos ponderados de las palabras a ser desambiguadas de la subjerarquía y el área total (número de conceptos) de la subjerarquía por debajo de c. La *nhyp* se calcula para cada concepto en WordNet, de forma que se satisfaga la ecuación 2.17, la cual expresa la relación entre la altura, el número de hipónimos de cada sentido y el número total de sentidos en una subjerarquía suponiéndola homogénea y regular.

La evaluación del método se realizó sobre parte del SemCor, obteniendo unos resultados prometedores (61,4%), además se tuvieron que adaptar los métodos de Yarowsky (1992) y Sussna (1993) para poder compararlos. Los resultados obtenidos fueron superiores que los métodos citados anteriormente.

Fernández-Amorós et al. (2001a) han investigado diferentes formas de utilizar la medida distancia conceptual en WSD, a partir de la medida de distancia conceptual propuesta por Agirre y Rigau (1996). Para ello introdujeron algunos parámetros e incorporaron algunas relaciones semánticas con el objeto de refinar y realizar una evaluación muy exhaustiva para cada una de las combinaciones propuestas. Para realizar la evaluación exhaustiva probaron al sistema en más de 50 configuraciones distintas contra todos los nombres de la colección de textos del SemCor. Después de realizar todas las combinaciones y pruebas se comprobó que este algoritmo mejora en un 42% al de Agirre y Rigau (1996) y un 14% al de Lesk (1986) basado en ocurrencias en las definiciones de un diccionario.

Resnik (1995a; 1995b) exploró la medida de similitud semántica (muy relacionada con la distancia conceptual) de los significados de dos palabras. Para ello tomó como referencia la red jerarquía estructurada de WordNet y en concreto la relación *IS-A* para nombres. Él calculó el contenido informativo (*information content*) compartido de las palabras. Esto es una medida para la especificidad del concepto que subsume las palabras en la jerarquía *IS-A* de WordNet. Es decir, la idea que utiliza es que cuanto más específico sea el concepto que subsume dos o más palabras, más relacionados semánticamente se supone que están. Por ejemplo, los conceptos *nickel#2* y *dime#1* son ambos subsumidos por el concepto *coin#1* en la taxonomía de WordNet, mientras que el concepto más específico para *nickel#2* y *credit_card#1* que comparten es *medium_of_exchange#1*. Esto conlleva una dificultad que es medir la especificidad de un concepto respecto a otros. El hecho de contar únicamente con las relaciones IS-A en la taxonomía puede resultar engañoso, ya que una relación simple puede representar una granularidad fina en una parte de la taxonomía (por ejemplo, *zebra#1 IS-A equine#1*) y una no tan fina (por ejemplo, *carcinogen#1 IS-A substance#1*). Pero si se decide contar con otras relaciones de la taxonomía el problema puede agravarse más, por eso una alternativa para medir la especificidad es considerar el contenido informativo de un concepto (conocido en inglés *information content*).

Resnik realizó una experimentación con la jerarquía de WordNet y y el contenido informativo (*information content*) definido en la fórmula 2.18, la cual mide la similitud de dos conceptos.

$$Sim(w1, w2) = \max_{C \in subsumer(w1,w2)} [-\log Pr(c)] \qquad (2.18)$$

Donde *subsumer(w1,w2)* es el conjunto de *synsets* antecesores tanto a *w1* como a *w2* para cualquier sentido de ambas palabras. El concepto C que maximice la expresión en 2.18 será el más informativo que incluya o contenga a *w1* y *w2*. Por ejemplo, el concepto *nickel#2* (en el sentido de *coin*) y *mortgage#1* tienen en común solo al super-concepto *possession#2*, con una similitud de 8,17; el concepto *nickel#2* y *dime#1* tienen en común todos los

super-conceptos listados en la Tabla 2.6, siendo el más informativo o específico aquel que obtiene el valor de similitud de 13,51.

CONCEPTO C	SIMILITUD
< *coin*, 3566679 >	13,51
< *coin*, 3566477 >	12,52
< *cash*, 3566144 >	12,45
< *currency*, 3565780 >	11,69
< *money*, 3565439 >	11,27
< *tender*, 3562119 >	11,27
< *medium_of_exchange*, 3561702 >	11,21
< *asset*, 3552852 >	9,71
< *possession*, 11572 >	8,17

Tabla 2.6. Super-conceptos de < *nickel* > y < *dime* >.

Hay que forzar a que la probabilidad del concepto no pueda ser decreciente a medida que se suba en la taxonomía, por ejemplo si se tiene la relación c_1 IS-A c_2 entonces $Pr(c_1) < Pr(c_2)$. Con esto se garantiza que lo más abstracto, es decir en niveles superiores de la taxonomía, signifique lo menos informativo.

La probabilidad $Pr(c)$ se obtiene a partir de un corpus mediante el cálculo de la fórmula 2.19.

$$Freq(c) = \sum_{n \in words(c)} count(n) \tag{2.19}$$

Donde *words(c)* es el conjunto de nombres que tienen un significado incluido en el concepto c, es decir están subsumidos por el concepto c. Por lo tanto la probabilidad simplemente se calcula como la frecuencia relativa aplicando la fórmula 2.20.

$$Pr(c) = \frac{Freq(c)}{N} \tag{2.20}$$

Donde N es el número total de nombres observados. Se asume que existe un único nodo raíz virtual, cuyos hijos son los nodos raíces originales. Por lo tanto, si dos significados tienen el nodo raíz virtual como su único límite superior, su valor de similitud es cero.

La Tabla 2.7 muestra los valores de la similitud semántica para varios pares de nombres, e indica al concepto más informativo que los incluye.

Palabra 1	Palabra 2	Similitud	Concepto más Informativo
doctor	nurse	9.4823	(health professional)
doctor	lawyer	7.2240	(professional person)
doctor	man	2.9683	(person,individual)
doctor	medicine	1.0105	(entity)
doctor	hospital	1.0105	(entity)
doctor	health	0.0	virtual root
doctor	sickness	0.0	virtual root

Tabla 2.7. Valores de similitud de pares de nombres.

Otro trabajo es el de Mihalcea y Moldovan (1999a) que presenta un método para desambiguar nombres, verbos, adjetivos y adverbios en un texto, utilizando WordNet y basándose en la *densidad semántica* entre palabras. Hay que aclarar que la desambiguación se hace en el contexto de WordNet. En este caso, la densidad semántica se mide por el número de palabras comunes que están dentro de una distancia semántica de dos o más palabras. Es decir, cuanto más cerca sea la relación semántica entre dos palabras mayor será la densidad semántica entre ellas. La densidad semántica funciona bien en los MRD´s uniformes. En realidad hay algunas lagunas en la representación del conocimiento y la densidad semántica sólo puede aportar una estimación de la relaciones semánticas entre palabras. Pero tiene la ventaja de que es muy fácil de medirla en un MRD como WordNet, por tal motivo se utilizó en este método.

El método utiliza Internet como un inmenso corpus, para reunir información estadística y clasificar los sentidos de las palabras, y WordNet, para medir la densidad semántica. También, este método consta de dos algoritmos; el primero, filtra los sentidos de los nombres y elige los dos mejores sentidos de todos los posibles, y el segundo, toma la salida del primero y devuelve el par de palabras con los sentidos anotados. A continuación se de-

tallarán los dos algoritmos para realizar la desambiguación de un par de palabras verbo-nombre.

Algoritmo 1 : *Filtrar los sentidos posibles del nombre.*
Este algoritmo clasifica los sentidos de los nombres con el objetivo de mejorar la precisión en el cálculo de la densidad conceptual entre un verbo y un nombre. La técnica utilizada en esta tarea es un método estadístico sobre todos los textos accesibles por internet. La entrada a este algoritmo son pares Verbo-Nombre y manteniendo al verbo constante. Posteriormente se aplican los siguientes pasos para clasificar los sentidos de los nombres:

Paso 1: Se hace una lista de similitud para cada sentido del nombre. Por ejemplo en 2.21, se muestra la lista para el nombre N que tiene m sentidos:

$$\begin{array}{l}(N^1,\ N^{1(1)},\ N^{1(2)},\ldots,N^{1(k1)})\\(N^2,\ N^{2(1)},\ N^{2(2)},\ldots,N^{2(k2)})\\\quad\vdots\\(N^m,\ N^{m(1)},\ N^{m(2)},\ldots,N^{m(km)})\end{array} \tag{2.21}$$

Donde $N^{i(s)}$ representa el número sinónimo s del sentido N^i del nombre N, según lo define la estructura de WordNet. Un ejemplo de lista de similitud para los dos primeros sentidos de la palabra *report*, se muestra en 2.22.

$$\begin{array}{l}(report, study)\\(report, newsreport, story, account, writeup)\end{array} \tag{2.22}$$

Paso 2: Se forman pares Verbo-Nombre, manteniendo la palabra verbo constante según se muestra en 2.23:

$$\begin{array}{l}(V-N^1,\ V-N^{1(1)},\ldots,V-N^{1(k1)})\\(V-N^2,\ V-N^{2(1)},\ldots,V-N^{2(k2)})\\\quad\vdots\\(V-N^m, V-N^{m(1)},\ldots,V-N^{m(km)})\end{array} \tag{2.23}$$

Un ejemplo para el verbo *investigate* y las palabras de la lista de similitud del nombre *report*, se muestra en 2.24.

$$\begin{array}{l}(investigate - report, investigate - study) \\ (investigate - report, investigate - news\_report, \\ investigate - story, investigate - account, \\ investigate - write\_up)\end{array} \quad (2.24)$$

Paso 3: En este paso se busca en Internet y se clasifican los sentidos. Para ello se realiza una búsqueda de cada conjunto de pares definido en 2.23, utilizando el buscador Altavista. Este nos devuelve la frecuencia de ocurrencias de la palabra V para el sentido de la palabra N. Es decir, usando estas consultas se obtiene el número de aciertos para cada sentido i del nombre N y esto nos da una clasificación de los m sentidos del nombre y como se relacionan con el verbo V. Por ejemplo, lo que se debe buscar en Altavista[11] para el verbo *investigate* y la primera entrada de la lista de similitud del nombre *report* será lo siguiente; ("investigate report" OR "investigate study"). Para este caso, Altavista devolvió 478 ocurrencias, y para el otro ("investigate-report", "investigate-news_report", "investigate-story", ´´investigate-account", "investigate-write_up") 281 ocurrencias.

Algoritmo 2 : *Determinar la densidad conceptual entre dos palabras (verbo y nombre).*

La densidad conceptual es difícil de calcular para palabras de distinta categoría léxica, por ejemplo entre verbos y nombres, sin un gran corpus o un MRD que tenga relaciones semánticas entre distintas categorías léxicas. Como esos recursos no se tienen actualmente, para conseguir tales relaciones semánticas se considerarán las definiciones o glosas de cada concepto en WordNet, ya que aportan un mini-contexto de ese concepto. En este algoritmo se considera que ese mini-contexto de las glosas de WordNet, es un recurso de información lingüística

[11] `http://www.altavista.com`

suficientemente rico para calcular la densidad conceptual entre palabras.

La entrada de información a este algoritmo es el verbo a desambiguar y los dos primeros sentidos del nombre previamente clasificados por el algoritmo 1. A continuación, determina el sentido para el verbo y el nombre, mediante el cálculo previo de la densidad conceptual para cada par $V_i - N_j$. Este cálculo se realiza mediante los pasos siguientes:

Paso 1: Se extraen todas las glosas de V_i incluidas en su jerarquía, inclusive la suya. La razón de esta decisión se debe a que en WordNet una glosa explica o define a un concepto y además aporta ejemplos típicos de ese concepto. Con objeto de determinar las jerarquías más apropiadas para nombres y verbos, se realizaron unos experimentos usando SemCor y se llegó a la conclusión de que la jerarquía de nombres debería incluir todos los nombres en la clase de N_j. Y la jerarquía del verbo V_i se toma como la jerarquía del más alto hipónimo h_i del verbo V_i. Como es necesario considerar una jerarquía grande entonces esto lo proporcionan los sinónimos y los hipónimos directos. Además, al cambiar el rol del corpus por las glosas, se obtendrán mejores resultados si más glosas son consideradas.

Paso 2: Se seleccionan los nombres de estas glosas, formando el mini-contexto del verbo. A cada uno de estos nombres se le asocia un peso, el cual indica el nivel en la subjerarquía del verbo en cuya glosa el nombre se encontró.

Paso 3: Se determinan los nombres a partir de la subjerarquía del nombre, inclusive N_i.

Paso 4: Se calcula la densidad conceptual CD_{ij} de los conceptos comunes entre los nombres obtenidos en el paso 2 y los obtenidos en el paso 3 aplicando la fórmula 2.25.

$$CD_{ij} = \frac{\sum_k^{|cc_{ij}|} W_k}{\log(descendientes_j)} \tag{2.25}$$

Donde $|cc_{ij}|$ es el número de conceptos comunes entre las jerarquías de V_i y N_j. W_k son los niveles de los nombres en

la jerarquía del verbo V_i. Y $descendientes_j$ es el número total de palabras en la jerarquía del nombre N_j.
Como los nombres con una gran jerarquía tienden a tener un gran valor para $|cc_{ij}|$, la suma de los valores para los conceptos comunes se normaliza con respecto a la dimensión de la jerarquía del nombre. Así el tamaño de una jerarquía crece exponencialmente con su profundidad, por eso se utiliza el logaritmo del número total de descendientes en la jerarquía $(log(descendientes_j))$.

Paso 5: La combinación idónea entre los sentidos del verbo y el nombre $V_i - N_j$ son los que obtienen el más alto valor de CD_{ij}.

Este método se evaluó con 384 pares de palabras seleccionadas a partir de los dos primeros ficheros del SemCor (fichero br−a01 y br−a02). Los pares de palabras obtenidos para Verbo-Nombre fueron 200, Adjetivo-Nombre 127 pares y Adverbio-Verbo 57 pares. Los resultados obtenidos al aplicar el método de desambiguación fueron de 86,5% para nombres, 67% para verbos, 79,8% para adjetivos y 87% para adverbios.

Magnini y Strapparava (2000) han estudiado la influencia de los dominios en la desambiguación de los sentidos. Así, para cada palabra en un texto se elige una etiqueta de un dominio en vez de la etiqueta del sentido. Por ejemplo, dominios como "medicina" o "deporte" describen de una manera natural las relaciones semánticas entre los sentidos de las palabras, además de reunirlos en grupos homogéneos. Una consecuencia de aplicar los dominios obtenidos de WordNet es que se reduce la polisemia de las palabras, ya que el número de dominios para una palabra es generalmente más bajo que el número de sentidos para la misma. Por ejemplo, la palabra *book* tiene siete sentidos diferentes en WordNet 1.6, y tres de ellos (*book#1, book#2 y book#7*) se pueden agrupar bajo el mismo dominio de *"publishing"*, por lo tanto se reduce su polisemia de 7 a 5 sentidos. A esta variante de WSD se le conoce como *Word Domain Disambiguation* (WDD). Para realizar este tipo de desambiguación se presentaron dos algorit-

mos, los cuales utilizan dos alternativas distintas para medir la frecuencia del dominio.

Algoritmo 1 : *Frecuencia del dominio según el texto.*
Este tipo de algoritmo utiliza dos pasos. El primero consiste en considerar todas las palabras del texto y en incrementar el contador del dominio en uno cada vez que una de esas palabras pertenezca a ese dominio. En segundo lugar, se consideran todas las palabras y aquel dominio con puntuación más alta será el elegido como el resultado de la desambiguación.

Algoritmo 2 : *Frecuencia del dominio según las palabras.*
En esta versión, la frecuencia de cada dominio perteneciente a la palabra se obtiene mediante el cociente de los sentidos pertenecientes al dominio, entre el total de sentidos de la palabra. Por ejemplo, si la palabra es *book*, el dominio *publishing* recibirá un valor de 0,42 (3 sentidos pertenecientes a *publishing* de 7 totales), mientras que los otros dominios recibirán cada uno un valor de 0,14 (un único sentido para cada dominio de 7 totales).

Los resultados obtenidos de los distintos experimentos realizados fueron bastante esperanzadores, debido a que se obtuvieron unos porcentajes bastantes elevados de precisión. Los resultados son de 85% para el primer algoritmo y 86% para el segundo.

2.4 Métodos mixtos

Últimamente, se han realizado muchos experimentos de WSD usando de forma combinada varias fuentes de conocimiento léxico (estructurado y no estructurado), así como diferentes técnicas para explotar dicho conocimiento. Esta técnicas difieren en:

1. El tipo de recurso léxico utilizado en los diferentes pasos del método propuesto (corpora, MRD´s, tesauros, bases de conocimiento léxicas).
2. Las características particulares de esos recursos que se utilizan durante la tarea de desambiguación.

3. Las medidas utilizadas para comparar similitudes entre unidades léxicas.

2.4.1 Mixto Corpus y tesauro

El trabajo de Yarowsky (1992) usó el *Roget´s Thesaurus* para dividir a *Grolier´s Encyclopedia* y reunir a las palabras por cada categoría. En este caso, la tarea de WSD en vez de realizarla a nivel de sentido se realiza a nivel de categoría de Roget (las palabras se dividen en 1042 categorías semánticas). Este método obtuvo una media de desambiguación correcta de alrededor del 92% en 12 palabras polisémicas.

Con un enfoque similar, los autores Liddy y Paik (Liddy y Paik, 1993) usaron los códigos semánticos por materias de LDOCE y el corpus *Wall Street Journal* para calcular una matriz de correlación de los códigos de materia. Para 166 oraciones anotadas con POS obtuvieron un porcentaje de acierto del 89% en la asignación de los códigos de materias (las palabras se dividen en 122 categorías semánticas).

2.4.2 Mixto WordNet y Corpus

Ribas (1995a) presentó un conjunto de técnicas y algoritmos para desambiguar nombres que acompañaban a un verbo, formando los dos un complemento. Para ello aplicó restricciones seleccionales sobre clases de WordNet y el sistema lo entrenó con un corpus no supervisado.

Resnik (1995a; 1995b) exploró la medida de similitud semántica (muy relacionada con la distancia conceptual) de los significados de dos palabras. Para ello tomó como referencia la red jerarquía estructurada de WordNet y en concreto la relación *IS-A* para nombres. Él calculó el contenido informativo (*information content*) compartido de las palabras. Esto es una medida para la especificidad del concepto que subsume las palabras en la jerarquía *IS-A* de WordNet. Es decir, la idea que utiliza es que cuanto más específico sea el concepto que subsume dos o más palabras, más relacionados semánticamente se supone que están.

2.4.3 Mixto métodos MRD´s

Wilks y Stevenson (1997) propusieron un sistema de desambiguación que utilizaba varios *taggers* parciales, usando cada uno recursos de conocimiento independientes. Ninguno de estos *taggers* desambiguan totalmente los textos, sino que cada uno proporciona tanta información como sea posible, y sus salidas se combinarán para realizar la desambiguación final. Los *taggers* comentados anteriormente incorporan como recursos de información: *Part-Of-Speech (POS)*, definiciones de diccionarios y etiquetas de dominios (categorías de thesaurus). Finalmente, utilizan un mecanismo que combina los resultados de los procesos que utilizan cada una de los recursos de información mencionados. Los resultados de desambiguación de las palabras polisémicas alcanzaban un 88%.

Rigau *et al.* (1997) presentaron un método que puede utilizarse para desambiguar el sentido de las palabras en un corpus no-etiquetado del castellano. Combinaron distintas técnicas heterogéneas y recursos léxicos independientes, obteniendo una precisión entre el 79% para polisémicas y el 83% para todas. Este método no-supervisado utiliza información de las definiciones de los diccionarios MRD´s para construir vectores, que representan conocimiento, probar distintas técnicas y medir la similitud con el objetivo de asignar el sentido correcto. También, se utiliza un MRD bilingüe para asignar categorías semánticas de WordNet a las palabras, además de realizar un proceso de entrenamiento no-supervisado con objeto de seleccionar palabras relevantes para cada categoría semántica. En este trabajo también se aplicó la distancia conceptual a los conceptos de WordNet.

2.4.4 Mixto relaciones sintácticas y corpus

Stetina *et al.* (1998) presentaron un método híbrido que se puede clasificar, como un método WSD basado en corpus supervisados y que utiliza relaciones sintácticas para obtener el sentido de las palabras. Además, también resuelve considerablemente el problema de la poca densidad de datos ya que con un corpus de entrenamiento muy pequeño obtiene una precisión media del 80,3%. El principal objetivo del método es asignar el sentido apropiado

en WordNet a todas las palabras *content-word* de cualquier frase sintácticamente analizada. Los sentidos de las palabras son determinados por la combinación más probable en todas la relaciones sintácticas derivadas de la estructura gramatical de la frase. Para ello el método utiliza las llamadas relaciones sintáctico-semánticas extraídas del análisis sintáctico.

El método de Stetina (1998) desambigua el sentido de las palabras cuya categoría gramatical este contenida en la base de datos léxica de WordNet. Para ello, requiere de un corpus de entrenamiento anotado sintáctica y semánticamente y preferiblemente supervisado. Este método obtiene del corpus un conjunto de relaciones sintáctico-semánticas (*head-modifier*). Por un lado se obtienen una serie de relaciones categoriales y por otro un conjunto de relaciones semánticas, ambas a partir del corpus de entrenamiento.

2.4.5 Mixto WordNet e Internet como corpus

Otro trabajo es el de Mihalcea y Moldovan (1999a) que presenta un método para desambiguar nombres, verbos, adjetivos y adverbios en un texto, utilizando WordNet y basándose en la *densidad semántica* entre palabras. El método utiliza Internet como un inmenso corpus, para reunir información estadística y clasificar los sentidos de las palabras, y WordNet, para medir la densidad semántica. También, este método consta de dos algoritmos; el primero, filtra los sentidos de los nombres y elige los dos mejores sentidos de todos los posibles, y el segundo, toma la salida del primero y devuelve el par verbo-nombre con los sentidos anotados.

2.5 Clasificación alternativa

Esta sección presenta otra clasificación de los sistemas WSD, desde el punto de vista de si requieren ejemplos etiquetados manualmente o no. Así, los sistemas supervisados se definen como aquellos sistemas que necesitan ejemplos con las palabras a desambiguar etiquetadas con su sentido correcto para posteriormente entrenarse. Sin embargo, los sistemas no-supervisados se definen

como aquellos sistemas que no necesitan estos ejemplos para realizar la desambiguación del sentido de las palabras.

La gran mayoría de métodos presentados en las secciones anteriores utilizan, para desambiguar, un recurso léxico básico, un corpus de entrenamiento, un conjunto de reglas y/o colocaciones de palabras en una oración, etc. Conseguir esta información aunque sea en pequeñas cantidades no es posible muchas veces. En particular, cuando se necesita información de dominios especializados, ya que estos recursos léxicos muchas veces no están disponibles. Por ejemplo, los sistemas de recuperación de información deben ser capaces de manejar colocaciones de palabras en textos para un dominio determinado. Los diccionarios generales no son útiles para tratar las colocaciones de un dominio específico.

Para resolver este problema, se han utilizado los métodos no supervisados, ya que obtienen los datos de entrenamiento semiautomáticamente. Estrictamente hablando, la desambiguación no supervisada no es posible si no se realiza una anotación del sentido a los textos. Por lo tanto debe haber un algoritmo que anote el sentido correspondiente a las palabras de los textos y para ello se requiere conocer la caracterización de cada uno de los sentidos. Sin embargo, la distinción de los sentidos (*sense discrimination*) se puede realizar de un modo completamente no supervisado, ya que se pueden agrupar las palabras ambiguas en grupos de contextos y realizar la distinción entre esos grupos sin etiquetarlos.

Uno de los métodos no supervisados es el propuesto por Yarowsky (1992), el cual proyecta la polisemia sobre las categorías definidas en el tesauro Roget, y entrenó a modelos estadísticos a categorías en vez de a palabras individuales. Por consiguiente, se observó en ese trabajo que las palabras monosémicas asociadas a cada categoría suministraba co-ocurrencias estadísticas fiables y el ruido estadístico presentado por las palabras polisémicas era tolerable sin ninguna anotación supervisada de los sentidos.

Otro método no supervisado es el propuesto por Schütze (1998) denominado *context-group discrimination*, el cual demuestra que la tarea de desambiguar los sentidos de las palabras es fundamental para mejorar los sistemas de recuperación de información. El algoritmo de desambiguación se basa en la agrupación de ocurren-

cias de una palabra ambigua, y los grupos obtenidos consisten en ocurrencias contextualmente similares. Los sentidos se interpretan como grupos de contextos similares de la palabra ambigua. Las palabras, contextos y sentidos se representan en un espacio vectorial de palabras multidimensional. Así, en vez de formar una representación del contexto a partir de las palabras en que la palabra ambigua aparece directamente con un contexto particular, se forma la representación del contexto a partir de las palabras en que estas palabras aparecen unas con otras en el corpus de entrenamiento. El algoritmo es automático y no supervisado tanto en el entrenamiento como en la aplicación: los sentidos se inducen a partir de un corpus sin ejemplos de entrenamiento etiquetados u otro recurso de conocimiento externo.

Otros investigadores como Justeson y Katz (1995) proponen un método para adquirir automáticamente ejemplos de entrenamiento para desambiguar el sentido de los adjetivos.

El trabajo de Yarowsky (1995) excluye completamente la supervisión manual mediante la adquisición automática de los datos de entrenamiento a partir de un diccionario. Además, Yarowsky utilizó restricciones en el discurso para descartar el ruido a partir de las listas de decisión. Los resultados del experimento de Yarowsky demuestran que la realización de este método es equivalente a la conseguida por los métodos de aprendizaje supervisados.

El trabajo de Mihalcea y Moldovan (1999a) utiliza Internet como un inmenso corpus, para reunir información estadística y clasificar los sentidos de las palabras, y WordNet, para medir la densidad semántica. También, este método consta de dos algoritmos; el primero, filtra los sentidos de los nombres y elige los dos mejores sentidos de todos los posibles, y el segundo, toma la salida del primero y devuelve el par verbo-nombre con los sentidos anotados.

El trabajo de Leacock *et al.* (1998) demuestra cómo técnicas basadas en el conocimiento se pueden usar para localizar automáticamente corpus de entrenamiento. Describe un clasificador estadístico que usa contexto general, contexto local o la combinación de los dos para identificar el sentido de las palabras. Las relaciones léxicas de WordNet se usan como base de conocimiento

para localizar automáticamente ejemplos de entrenamiento en un corpus de ámbito general.

2.5.1 Clasificación de métodos según Senseval

En esta sección se presentarán los sistemas de WSD más recientes presentados a SENSEVAL-1 y SENSEVAL-2. SENSEVAL es una competición científica sobre WSD (Kilgarriff, 1998) que se han celebrado al estilo de la agencia norteamericana ARPA o MUC. Actualmente hay muchos programas para determinar automáticamente el sentido de una palabra en un contexto. El propósito de SENSEVAL es evaluar la potencia y la debilidad de tales programas con respecto a diferentes palabras, diferentes variedades de lengua y diferentes lenguas. El primer SENSEVAL, denominado SENSEVAL-1, se realizó en el verano de 1998 para el inglés, francés e italiano. El segundo SENSEVAL, denominado SENSEVAL-2, se realizó en Julio de 2001 sobre 12 lenguajes: checo, holandés, inglés, estonio, vasco, chino, danés, italiano, japonés, coreano, español y suizo.

Estos sistemas que se van a presentar difieren tanto en la metodología utilizada como en la entrada de datos requerida, por lo tanto es muy difícil hacer una comparación aceptable. Para realizar esta comparación, en esta Tesis se sigue la misma clasificación utilizada en SENSEVAL, es decir los sistemas se clasificarán en dos tipos: sistemas supervisados y sistemas no-supervisados. La Tabla 2.8 muestra la clasificación de los sistemas supervisados, y la Tabla 2.9 la de los sistemas no-supervisados, que se presentaron a la competición SENSEVAL-1 para realizar la tarea de *English Lexical Sample*. La cual consiste en evaluar un conjunto de palabras seleccionadas previamente (nombres, verbos y adjetivos) de los ejemplos del texto. La descripción de cada uno de estos sistemas se presenta en el trabajo de Kilgarrif y Palmer (2000), aunque algunos son presentados también en artículos personales de los autores[12].

[12] Toda la información que se va a describir en esta sección se puede encontrar con mucho mayor detalle en la dirección `http://www.itri.bton.ac.uk/events/senseval/`.

Sistemas supervisados	Nombre	Contacto
Bertin, U Avignon	Avignon	De Loupy
Educ Testing Service, Princenton	Ets-pu	Leacock
John Hopkins U	Hopkins	Yarowsky
Korea U	Korea	Ho Lee
UNC Asheville	Grling-sdm	O'Hara
Tech U Catalonia, Basque U	upc-ehu-su	Agirre
U Durham	Durham	Hawkins
U Manitoba	Manitoba-ks	Suderman
U Manitoba	Manitoba-dl	Lin
U Tilburg	Tilburg	Daelemans

Tabla 2.8. Sistemas supervisados participantes en SENSEVAL-1

Sistemas No-supervisados	Nombre	Contacto
CL Research, USA	Clres	Litkowski
Tech U Catalonia, Basque U	upc-ehu-un	Agirre
U Ottawa	Ottawa	Barker
U Manitoba	Mani-dl-dict	Lin
U Sunderland	Suss	Ellman
U Sussex	Sussex	McCarthy
U Sains Malaysia	Malaysia	Guo
Xerox-Grenobe, CELI-Torino	Xeroxceli	Segond
CUP/Cambridge Lang Services	Cup-cls	Harley

Tabla 2.9. Sistemas no-supervisados participantes en SENSEVAL-1

Todavía más recientes son los sistemas presentados a SENSEVAL-2, realizado en el verano del 2001. En esta última competición los sistemas WSD se han evaluado con tres tipos de tareas sobre 12 lenguas diferentes. En la tarea *"all-words"*, la evaluación se realiza sobre todas las palabras de los ejemplos del texto. En la tarea *"lexical-sample"*, la evaluación se realiza sobre una palabra concreta de los ejemplos del texto. En la tarea *"translation task"*, la distinción de los sentidos se corresponden con la traducción de una palabra en otro lenguaje. Las tareas fueron las siguientes: *"All-words"*en checo, holandés, inglés y estonio. *"Lexical-sample"* en vasco, chino, danés, inglés, italiano, japonés, coreano, español y sueco. *"Translation"* en japonés. En total se presentaron a esta competición 90 sistemas.

Una vez descritas las principales características de la competición SENSEVAL-2, a continuación y siguiendo su propia clasificación de sistemas WSD tanto supervisados como no-supervisados, se presentarán en las dos siguientes secciones los sistemas más relevantes en cada una de las tareas *"all-words"* y *"lexical-sample"*. Consideramos como más relevantes aquellos sistemas que han obtenido mejores resultados en la competición.

2.5.2 Descripción de los sistemas más relevantes para la tarea *English-all-words*

La Tabla 2.10 muestra la clasificación de los sistemas supervisados, y la Tabla 2.11 la de los sistemas no-supervisados, que se presentaron a la competición SENSEVAL-2 para realizar la tarea de *English all-words*.

Sistemas supervisados	Nombre	contacto
University Basque Country	ehu-dlist-all	Agirre & Martinez
University of California	University_California	Chao
Lab. Informatique d'Avignon	LIA-Sinequa_AllWords	Crestan et al.
University of Antwerp	ANTWERP	Hoste
Southern Methodist Univ.	SMUaw	Mihalcea
University of Maryland	UMD-SST:	Resnik et al.

Tabla 2.10. Sistemas supervisados participantes en SENSEVAL-2 en la tarea *English all-words*

Como se ha comentado anteriormente, el criterio de selección escogido para describir este tipo de sistemas ha sido en base a los resultados obtenidos en la competición para cada una de las tareas. Así, a continuación se describirán los dos sistemas supervisados y no-supervisados más relevantes, es decir que hayan obtenido los mejores resultados en la competición. Los sistemas supervisados que se describirán son *SMUaw* y *Antwerp*, y los no-supervisados son *UNED-AW-U2* y *CL Research (DIMAP)*.

Sistemas supervisados. *El sistema SMUaw* utiliza para el proceso automático de entrenamiento los datos proporcionados por el SemCor. Por lo tanto se clasifica dentro de los sistemas supervisados. La desambiguación semántica de una palabra se realiza en

Sistemas No-supervisados	Nombre	contacto
University of Sussex	Sussex-sel	Carroll & McCarthy
University of Sussex	Sussex-sel-ospd	Carroll & McCarthy
University of Sussex	Sussex-sel-ospd-ana	Carroll et al.
University of Maryland	UMD-UST	Diab & Resnik
UNED University	UNED-AW-T	Fernandez-Amoros
UNED University	UNED-AW-U2	Fernandez-Amoros
Universiti Sains Malaysia	usm_english_tagger	Guo
Illinois Institute of Technology	IIT1, IIT2, IIT3	Haynes
CL Research	DIMAP	Litkowski
Instituto Trentino di Cultura	irst-eng-all	Magnini
University of Sheffield	University_Sheffield	Preiss

Tabla 2.11. Sistemas no-supervisados participantes en SENSEVAL-2 en la tarea *English all-words*

base a sus relaciones semánticas con las palabras que le preceden y que le siguen. Para ello, se creó un corpus muy grande de pares palabra-palabra anotadas con su sentido mediante el uso de:

1. Los ejemplos de las glosas de cada uno de los *synsets* de WordNet 1.7.
2. SemCor, el cual tuvo que tratarse para cambiar los sentidos anotados de la versión 1.6 del WordNet por los de la versión 1.7. En el caso de que se encontrara un sentido que no estaba en WordNet 1.6, se asignaba el sentido 0 por defecto. SemCor 1.7a está disponible en `http://www.seas.smu.edu/ rada/semcor`.
3. GenCor, el cual es un corpus alrededor de 160.000 palabras etiquetadas con sus sentidos. Este corpus se generó a partir de los ejemplos de SemCor y WordNet y para ello se utilizaron los principios de etiquetación de corpus presentados en el trabajo de Mihalcea y Moldovan (1999b). La metodología para crear este corpus se describe detalladamente en el trabajo de Mihalcea (2001).
   Un gran conjunto adicional de pares palabra-palabra generado a partir de los pasos 1 y 2. Para ello, se aplican un conjunto de heurísticas.

El *recall* de este algoritmo no es el 100 %, por lo tanto, se aplica una metodología para propagar los sentidos de las palabras

desambiguadas al resto de palabras ambiguas encontradas en el contexto. Y si aún siguen algunas palabras ambiguas entonces se le aplica el sentido más frecuente de WordNet. El algoritmo que utiliza este método se describe en el trabajo de Mihalcea y Moldovan (2000), aunque en una versión muy simple e inicial del algoritmo.

El sistema Antwerp también utiliza para el proceso automático de entrenamiento datos anotados previamente. Por lo tanto se clasifica dentro de los sistemas supervisados. Este método aplica técnicas de aprendizaje automático (*Machine learning*) para desambiguar automáticamente el sentido de las palabras en la tarea *English-all-words*. El sistema experto en palabras semánticas se entrena con el SemCor. El sistema experto en palabras semánticas combina diferentes tipos de algoritmos de aprendizaje como *memory based learning (TiMBL)* y *rule induction (Ripper)*, y toman diferentes fuentes de conocimiento como entrada:

1. La entrada de un algoritmo *memory-based learning (TiMBL)* es una vector de características, el cual está formado de la palabra objetivo con su lema, del sentido clasificado de la palabra a desambiguar y de tres palabras a derecha e izquierda junto con sus categorías gramaticales.
2. A un segundo algoritmo *memory-based learning (TiMBL)* se le introduce un vector de co-ocurrencias en el vector de características, el cual está formado por las palabras claves de una oración a la izquierda y a la derecha de la palabra a tratar. Este vector además contiene las palabras disponibles en las definiciones de WordNet.
3. La entrada al método *rule induction (Ripper)* es la información del contexto así como todas las posibles palabras claves en un contexto de tres oraciones. Ambos algoritmos *memory-based learning (TiMBL)* se validan para determinar el conjunto de parámetros óptimos para cada palabra semántica experta. Sobre los resultados de los clasificadores combinados y el sentido más frecuente de WordNet se realiza una votación. La arquitectura del sistema experto en palabras semánticas permite que los procesos de entrenamiento se ejecuten en paralelo. Así,

para clasificar un test de entrada dado, primero se comprueba si se dispone de una palabra semántica experta. Si es así, el algoritmo que mejor funcione sobre el conjunto de entrenamiento se aplica con el conjunto de parámetros óptimos con el objetivo de clasificar esa palabra. Si no es así, se devuelve al sentido más frecuente de WordNet.

Una descripción detallada de este algoritmo se presenta en el trabajo de Hoste *et al.* (2001).

Sistemas no-supervisados. *El sistema UNED-AW-T* es un sistema no-supervisado, ya que no necesitan ejemplos etiquetados para realizar la desambiguación del sentido de las palabras. Para realizar la desambiguación de todas las palabras se obtienen los lemas, se eliminan las palabras de parada del contexto y se detectan los nombres propios y los números. También se detectan las palabras compuestas que están en WordNet. A continuación, se eliminan mediante el fichero *cntlist* de WordNet los sentidos que no aparecen más del 10 % en los ficheros de WordNet.

También, se hace una matriz de pesos (similitud entre palabras) calculada a partir de los datos de 3200 libros de inglés obtenidos a partir de *Gutenberg Project* (`http://promo.net/pg/`). Estas similitudes son sensibles a las distancias entre las palabras en el corpus.

Otra segunda matriz denominada *relevance matrix* se construye mediante la medida de información mutua entre palabras (o etiquetas). Para la construcción del sistema WSD se utilizan diferentes heurísticas aplicadas en cascada como: heurística de expresiones monosémicas, filtros estadísticos, filtros de la *relevance matrix*, enriquecer el vector de características del sentido y heurísticas de finalización. Una descripción detallada de este algoritmo y de las heurísticas se presenta en el trabajo de Fernández-Amorós *et al.* (2001b).

El sistema CL Research (DIMAP) no necesita ejemplos para realizar la desambiguación del sentido de las palabras, ni procesos de entrenamiento, por lo tanto se clasifica dentro de los métodos no-supervisados. El sistema de desambiguación CL Research forma parte del software desarrollado para el diccionario electrónico

DIMAP. Este software se ha diseñado para utilizar cualquier diccionario como parte fundamental para realizar la desambiguación del sentido de las palabras. Los resultados para SENSEVAL-2 se generaron con WordNet y con *New Oxford Dictionary of English (NODE)*. Este proceso de desambiguación explota cualquier información que esté disponible en la base de datos léxica. El diseño de este sistema es similar al presentado en el SENSEVAL-1; sin embargo, muchos de los procedimientos de desambiguación no pudieron ser reimplementados por la forma en que se tenían que enviar los resultados. Para el SENSEVAL-2, se implementaron procedimientos especiales para examinar palabras compuestas y grupos contextuales (tanto con palabras específicas, Lesk utiliza la definición de las palabras, como con el análisis de categorías). Sin embargo no se emplearon restricciones sintácticas. Los resultados oficiales que se enviaron utilizaron solamente información disponible de WordNet. El diccionario NODE se utilizó para los datos de entrenamiento enviados por el SENSEVAL-2. Las definiciones de NODE se unieron automáticamente con WordNet, por lo tanto los resultados se pudieron comparar con los de WordNet para los datos de entrenamiento. Con este diseño del sistema se facilita el análisis de diferentes tipos de información, además de que con una implementación adicional se conseguirá una valoración de la importancia conseguida cuando se utilizan distinta información léxica.

2.5.3 Descripción de los sistemas para la tarea *English-lexical-sample*

La Tabla 2.12 muestra la clasificación de los sistemas supervisados, y la Tabla 2.13 la de los sistemas no-supervisados, que se presentaron a la competición SENSEVAL-2 para realizar la tarea de *English lexical-sample*.

A continuación se describirán los dos sistemas supervisados que hayan obtenido los mejores resultados en la competición SENSEVAL-2. Los sistemas supervisados que se describirán son *JHU-English* y *SMUls*.

Sistemas supervisados	Nombre	contacto
University Basque Country	ehu-dlist-all	Agirre & Martinez
University Basque Country	ehu-dlist-best	Agirre & Martinez
University of Sunderland	SUSS2	Canning et al.
Lab. Informatique d'Avignon	LIA-Sinequa_Lexsample	Crestan et al.
Tech University of Catalonia	TALP	Escudero et al.
UNED University	UNED-LS-T	Fernandez-Amoros
Instituto Trentino di Cultura	irst-eng-sample	Magnini
Stanford University	CS224N	Manning
Southern Methodist Univ.	SMUls	Mihalcea
University of Alicante	Univ._Alicante_System	Montoyo & Suarez
Univ. of Minnesota Duluth	Duluth1,..., DuluthC	Pedersen
University of Maryland	UMD-SST	Resnik et al.
Korea University	Kunlp	Seo, Lee & Rim
Johns Hopkins University	JHU-English	Yarowsky et al.

Tabla 2.12. Sistemas supervisados participantes en SENSEVAL-2 en la tarea *English lexical-sample*

Sistemas No-supervisados	Nombre	contacto
UNED University	UNED-LS-U	Fernandez-Amoros
Illinois Institute of Technology	IIT1, IIT2	Haynes
CL Research	DIMAP	Litkowski
ITRI (University of Brighton)	WASPS-Workbench	Tugwell

Tabla 2.13. Sistemas no-supervisados participantes en SENSEVAL-2 en la tarea *English lexical-sample*

El método presentado en esta Tesis, que forma parte del sistema completo que participó en esta tarea, se utilizó para desambiguar el sentido de los nombres. Este método se clasifica entre los no-supervisados, por lo tanto, se describirán todos los sistemas no-supervisados que participaron en la competición SENSEVAL-2. Los sistemas no-supervisados que se describirán son *UNED-LS-U*, *WASPS-Workbench*, *CL Research (DIMAP)* y *IIT1, IIT2.*

Sistemas supervisados. *El sistema JHU-English* para la tarea *lexical sample* consiste de 6 subsistemas de aprendizaje supervisados integrados. El resultado final se obtiene mediante la combinación de clasificadores. Los 6 subsistemas están formados por un método basado en listas de decisión (Yarowsky, 2000), un método basado en aprendizaje por transformación conducido por el error

(Brill, 1995; Florian y Ngai, 2001), un método basado en modelos de vectores mediante el coseno, método basado en decisiones y dos métodos basados en modelos probabilísticos de Naive-Bayes (uno entrenado sobre palabras y otro sobre lemas). Para cada subsistema las características que se incluyen no son solamente un conjunto de palabras en una ventana fija de contexto, si no que también incluyen una variedad de características sintácticas como sujetos, objetos directos, complementos circunstanciales y varias relaciones modificadoras de nombres y adjetivos. Estas relaciones se obtuvieron mediante la utilización de heurísticos basados en patrones sobre sintagmas nominales que están entre paréntesis en las oraciones (Florian y Ngai, 2001). Características adicionales como categorías gramaticales y lemas en cualquier posición sintáctica se extraen con el etiquetador *Brill-style POS* y se incluyen con el análisis morfológico de Yarowsky y Wicentokski (2000). La salida de cada subsistema se fusiona mediante un algoritmo combinado clasificador que usa votación, pesos y combinación de resultados.

El sistema SMUls utiliza cuatro pasos para desambiguar el sentido de las palabras en la tarea de *lexical-sample*.

1. Los datos son preprocesados aplicando los siguientes procesos: eliminación de las etiquetas SGML, la tokenización del texto y la identificación de las categorías gramaticales y los nombres de entidades.
2. Las palabras compuestas se identifican. Inicialmente se decide cual es la secuencia máxima que forma una palabra compuesta en WordNet. Los datos de entrenamiento y de test se dividen en base a como las palabras son etiquetadas. Por ejemplo, los ejemplos del contexto que tengan el verbo *dress down* se separan de los ejemplos que solo contiene al verbo *dress*. También se eliminan las palabras que son monosémicas así como los nombres propios.
3. Se extraen automáticamente patrones (mediante un conjunto de heurísticas) para cada palabra ambigua, mediante el uso de los ejemplos de WordNet (con los synsets de las palabras desambiguadas), de los ejemplos de SemCor y de los ejemplos de entrenamiento suministrados. Esos patrones se validan sobre

los datos de entrenamiento, y se guardan solamente aquellos que tienen una efectividad del 100 %. A continuación los patrones se aplican a los datos de test. Solo unos pocos ejemplos se pueden desambiguar de esta forma, pero con una muy alta confianza.

4. Este paso es el principal del sistema y desambigua todos los ejemplos que no se han desambiguado previamente. Para ello se usa un algoritmo de aprendizaje y un conjunto de características. El algoritmo de aprendizaje se entrena con los datos de entrenamiento proporcionados y a continuación se aplica a los datos de test. El algoritmo de aprendizaje utilizado es *Timbl* (Daelemans et al., 2001).

Sistemas no-supervisados. *El sistema UNED-LS-U* utilizado en la tarea de *English-lexical-sample* es el mismo que se utilizó para la tarea de *English-all-words*, que ya se describió anteriormente en la subsección de sistemas no-supervisados para la tarea de *English-all-words*. Una descripción detallada de este sistema se presenta en el trabajo de Fernández-Amorós *et al.* (2001b).

El sistema WASPS-Workbench integra lexicografía y WSD con el objetivo de beneficiarse de ambas cosas. El usuario introduce una palabra para ser analizada y el sistema calcula un *Word-Sketch*, que consiste en una página de significado estadístico con patrones de colocaciones para esa palabra introducida (que aparecen en el British National Corpus (BNC)). Sobre la base de estos patrones, el usuario prepara una lista de sentidos con orden y precisión y asigna sentidos a cada patrón particular. Esta primera asignación es usada iterativamente en un algoritmo de mejora progresiva (*bootstrapping*), el cual desambigua el corpus completamente. Los resultados para los lexicógrafos son un número determinado de *Word-Sketch*, que muestran patrones significativos para los sentidos individuales de la palabra. Sin embargo, para el proceso automático de WSD, se usa una lista de decisión, que consiste en patrones de relaciones gramaticales, palabras del contexto y n-gramas.

Para esta tarea de SENSEVAL-2 se asignan los sentidos a partir de una lista prefijada (en este caso WordNet).

El sistema CL Research (DIMAP) utilizado en la tarea de *English-lexical-sample* es el mismo que se utilizó para la tarea de *English-all-words*, que ya se describió anteriormente en la subsección de sistemas no-supervisados para la tarea de *English-all-words*.

El sistema IIT1, IIT2 para desambiguar una palabra marcada en la tarea *lexical-sample*, acumula todos los ejemplos de WordNet 1.7 (todo lo que va entre comillas) que se relacionan al *synset* propio de esa palabra y a cualquiera de sus *synsets* relacionados (todos los antecesores de las relaciones que sean padre, hijos inmediatos de las relaciones que sean hijo). Cada ejemplo debería tener una de sus palabras del *synset* o sus colocaciones como si fuera un mini corpus de los items del *lexical-sample* anotados. Es decir, se alinea el contexto de la palabra a desambiguar a la palabra *synset* del ejemplo.

Para cada palabra del ejemplo hay que encontrar la palabra en el contexto que se relacione mediante las relaciones de WordNet (dos palabras se consideran relacionadas si tienen un antecesor común con relaciones padre/hijo o pertenecen al mismo synset).

Se ponderan las veces que se combinan los ejemplos según:

1. proximidad léxica de casos que están cerca de una clase o categoría.
2. coincidencia exacta de las palabras o su categoría gramatical (POS) para palabras pertenecientes a una clase.
3. posiciones cambiadas de las palabras que coinciden.
4. ordenación de las palabras que coinciden.
5. y proximidad léxica del synset ejemplo para el sentido de la palabra candidata.

Finalmente, toda la información referente a la competición SENSEVAL-2 será publicada en las actas del Second International Workshop on Evaluating Word Sense Disambiguation Systems, que se realizó conjuntamente con el 39th Annual Meeting of the Association for Computational Linguistic ACL-2001. Además, actualmente todos los datos, resultados de la evaluación, descripción de tareas, descripción de sistemas y clasifi-

caciones de sistemas son de dominio público en la dirección `http://www.sle.sharp.co.uk/senseval2/`.

Volver al índice/Tornar a l'índex

3. Marcas de Especificidad como método de desambiguación léxica

En este capítulo se presenta el método de Marcas de Especificidad propuesto en esta Tesis para la desambiguación léxica de nombres en cualquier lengua[1]. Así en primer lugar presentaremos la arquitectura del sistema de Procesamiento del Lenguaje Natural utilizado junto con los recursos y herramientas lingüísticas utilizadas por el método. Y en segundo lugar se presenta detalladamente el método.

Una descripción breve sería: inicialmente la entrada al sistema de PLN está formada por un grupo de palabras, cuya categoría léxica es nombre, obtenidas a partir de una oración. Este grupo de palabras forma el contexto de entrada al sistema de desambiguación. Finalmente, el sistema PLN busca cada una de las palabras que forman el contexto en una base de conocimiento léxica, le aplica el método de desambiguación usando las Marcas de Especificidad y produce la salida de sus posibles sentidos.

Las principales aportaciones del método de Marcas de Especificidad se han presentado en los trabajos de Montoyo, Palomar y Rigau (2000a; 2000b; 2000; 2001; 2001).

El método de Marcas de Especificidad está actualmente definido para desambiguar las palabras cuya categoría léxica sea nombre, sin embargo el sistema queda abierto para trabajos futuros al tratamiento de otras categorías léxicas como verbos, adjetivos y adverbios.

[1] Este método funciona correctamente en cualquier lengua que tenga una base de conocimiento WordNet particular. Global WordNet Association es una organización pública y sin ánimo de lucro, que suministra una plataforma para discutir, compartir y relacionar WordNets para todos los lenguajes en el mundo. La dirección donde se detalla todo lo relacionado con esta organización es: *http : //www.hum.uva.nl/ ~ ewn/gwa.htm*

3.1 Arquitectura del sistema de PLN

En esta sección se describirá detalladamente la arquitectura utilizada en el desarrollo del sistema de PLN, que incluye el método de Marcas de Especificidad como núcleo principal de dicho sistema. La arquitectura del sistema de PLN propuesto se ilustra en la figura 3.1.

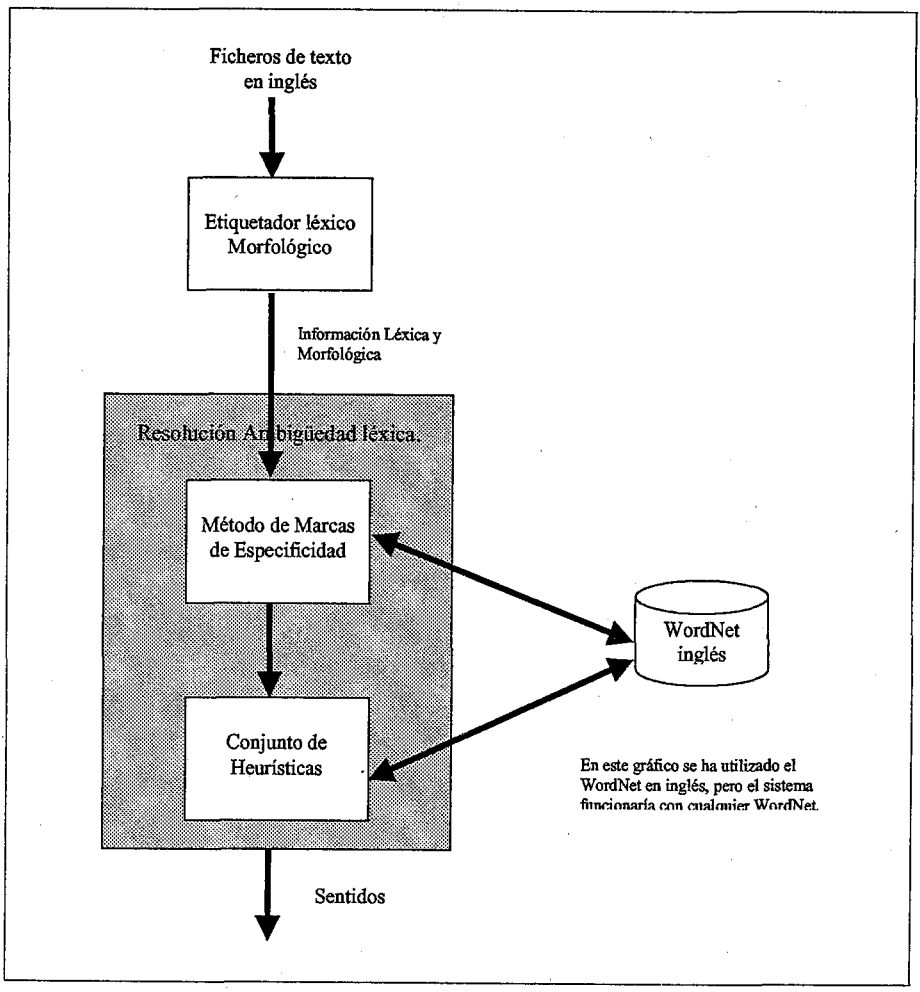

Figura 3.1. Sistema de PLN para WSD

A continuación se describen brevemente los recursos y herramientas propuestos por dicho sistema PLN.

- **Etiquetador léxico-morfológico:** genera las categorías gramaticales de cada palabra junto con su información morfológica, y proporciona al sistema independencia en la entrada de ficheros de texto o documentos.
- **WordNet:** proporciona información correspondiente a la taxonomía de nombres (Hiperonimia/hiponimia) así como todos los posibles sentidos de las palabras que forman el contexto en una oración de los textos de entrada.
- **Método de Marcas de Especificidad:** resuelve la ambigüedad léxica y se basa en el enfoque de cotejar el contexto de la palabra a desambiguar con información de un recurso de conocimiento externo. Por lo comentado anteriormente interesa tener un recurso que tenga las palabras y los conceptos organizados alrededor de clases (jerarquías), de tal forma que describan todas sus características semánticas. El método propuesto fue diseñado para explotar las relaciones jerárquicas de hiperonimia y hiponimia que proporciona la base de conocimiento léxica denominada WordNet.
- **Heurísticas:** utilizan la estructura de WordNet como fuente de información y devuelven el sentido de las palabras que no han podido desambiguarse por el método anterior. En total se han definido 6 heurísticas auxiliares asignando a cada una de ellas un nombre representativo de la acción que realizan, siendo la heurística del Hiperónimo, Definición, Hipónimo, Glosa Hiperónimo, Glosa Hipónimo y Marca Especificidad Común.

A continuación se presentará el etiquetador léxico morfológico utilizado con sus principales características y posteriormente el recurso léxico WordNet como fuente de información de los sentidos de las palabras.

3.2 Etiquetador léxico-morfológico

El sistema de PLN utilizado en esta Tesis funciona con cualquier etiquetador. La labor del etiquetador es transformar la secuencia

de caracteres de entrada en una secuencia de unidades significativas mediante el uso de reglas morfológicas y del diccionario. Es decir, anota a cada palabra que forma parte de una oración con una etiqueta[2] que indica como se usa esa palabra dentro de la oración. De hecho, la misma palabra puede ser un nombre en una oración y un verbo o un adjetivo en la siguiente. En el sistema de PLN propuesto se necesita un etiquetador para filtrar todas aquellas palabras cuya categoría léxica sea nombre a partir del texto de entrada. Para ello, se han utilizado dos etiquetadores diferentes, uno para textos en inglés y otro para textos en español.

El etiquetador léxico morfológico para textos en inglés utilizado en el sistema propuesto de PLN, se conoce con el nombre de *TreeTagger* (Schmid, 1994). Este etiquetador fue desarrollado dentro del Proyecto *"Textual corpora and tools for their exploration (TC)"* en el Instituto de Lingüística Computacional de la Universidad de Stuttgart. La característica fundamental del *TreeTagger* es analizar y desambiguar las categorías léxicas de las palabras para textos no restringidos en inglés. Para ello se basa en los modelos de Markov y en los árboles de decisión con el objetivo de obtener estimaciones más fiables.

En contraste a los etiquetadores basados únicamente en n-gramas, el etiquetador *TreeTagger* estima la transición de probabilidades a través de un árbol de decisión binario. La probabilidad de un trigrama dado se determina siguiendo su camino correspondiente a través del árbol de decisión hasta que se llega a una hoja. En la figura 3.2 se ilustra un ejemplo de un árbol de decisión utilizado por el *TreeTagger*.

Por ejemplo, si se quiere buscar la probabilidad de un nombre en la figura 3.2, el cual está precedido por un determinante y un adjetivo $p(NN \mid DET, ADJ)$. En primer lugar se debe responder a la pregunta del nodo raíz $tag_{-1} = ADJ?$. Como la etiqueta de la palabra previa es un ADJ, se sigue por el camino del árbol binario que responde con un ´´si ". A continuación se tiene la pregunta $tag_{-2} = DET?$, que también se cumple, por lo tanto se termina en el nodo hoja. Por último hay que buscar la probabilidad de la

[2] Las etiquetas utilizadas por estos son generalmente: nombres, verbos, pronombres, adjetivos, adverbios, preposiciones, interjecciones y conjunciones.

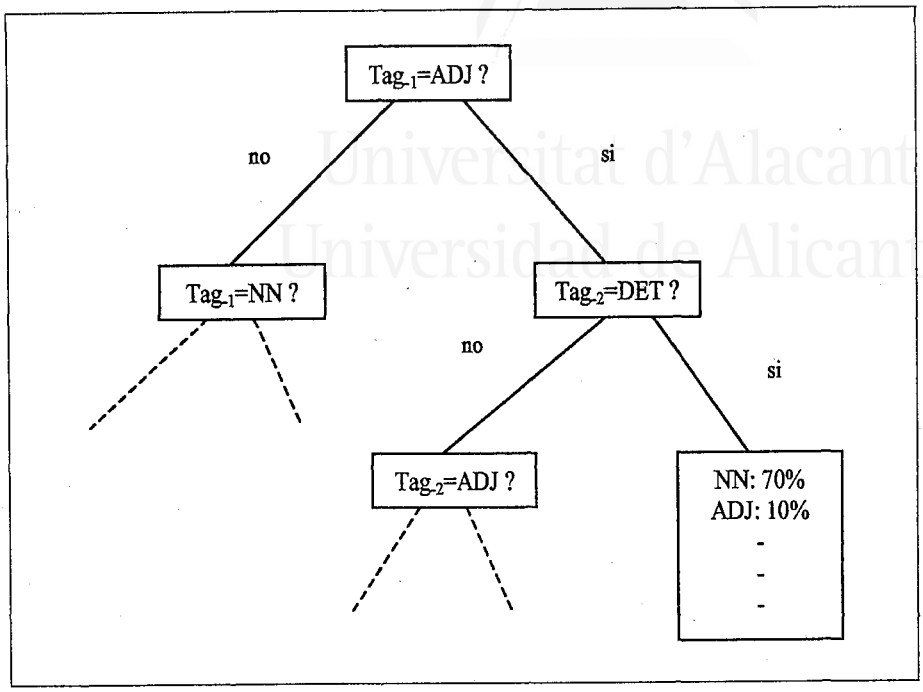

Figura 3.2. Un ejemplo de árbol de decisión

etiqueta *NN* en la tabla que está unida a ese nodo, en este caso 70 %. Los árboles de decisón son construidos recursivamente a partir de un corpus de entrenamiento (conjuntos de trigramas). Por ejemplo, este etiquetador se evaluó sobre los datos del Penn-Treebank obteniendo una efectividad del 96,36 %, frente a los 96,06 % que un etiquetador trigrama obtuvo sobre los mismos datos.

El *TreeTagger*, como todo método probabilístico, necesita una fase de entrenamiento previa a la fase de etiquetado. Una vez finalizada la fase de entrenamiento, el etiquetador léxico-morfológico ya se encuentra disponible para realizar el etiquetado de los textos de entrada. Para la fase de entrenamiento su autor partió del corpus etiquetado *Penn-Treebank*[3], y en concreto, empleó una porción del corpus de 2 millones de palabras. La salida de los textos de entrada, proporcionada por el etiquetador, es una relación

[3] Una relación de dichas etiquetas, así como una descripción detallada del proyecto se puede encontrar en la URL: *http* : *//www.cis.upenn.edu/* ~ *treebank*.

de palabras con sus lemas y una etiqueta que indica la categoría gramatical de la palabra junto con su información morfológica. El formato de salida tiene la estructura mostrada en la Tabla 3.1 para cada una de las oraciones de entrada del texto.

$palabra_1$	$etiqueta_1$	$lema_1$
$palabra_2$	$etiqueta_2$	$lema_2$
.....	.....	.....
$palabra_n$	$etiqueta_n$	$lema_n$

Tabla 3.1. Formato salida del etiquetador

Así, la Tabla 3.2 muestra un ejemplo de la salida del etiquetador TreeTagger[4] que consiste en la relación de palabras, que forman la oración, con sus lemas y una etiqueta que indica la categoría gramatical de la palabra junto con su información morfológica.

Palabra	*etiqueta*	*lema*
The	*DT*	*the*
TreeTagger	*NP*	*TreeTagger*
is	*VBZ*	*be*
easy	*JJ*	*easy*
to	*TO*	*to*
use	*VB*	*use*
.	*SENT*	.

Tabla 3.2. Salida del etiquetador para una oración ejemplo

El etiquetador léxico morfológico para textos en español utilizado en el sistema propuesto de PLN analiza y desambigua las categorías léxicas de las palabras para textos en español basándose en los modelos estocásticos ECGI Extendidos (Pla et al., 2000; Pla, 2000) desarrollado por el Grup de Processament del Llen-

[4] Una descripción detallada de este etiquetador se puede encontrar en la URL: `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`

guatge Natural de la Universidad Politècnica de València. Para la fase de entrenamiento sus autores partieron de un corpus etiquetado que fue revisado por un experto. Una vez que se ha finalizado la fase de entrenamiento, el etiquetador léxico morfológico ya se encuentra disponible para realizar etiquetados (fase de etiquetado). La fase de etiquetado o desambiguación léxica del texto usa el analizador morfológico MACO+ (Carmona et al., 1998) desarrollado en el Grup de Processament de la Llengua de la Universitat Politècnica de Catalunya. MACO+ segmenta el texto en tokens y proporciona todas las categorías gramaticales posibles para cada token junto con su información morfológica según el juego de etiquetas PAROLE (Martí et al., 1998). La salida de MACO+ y el modelo léxico entrenado constituyen las probabilidades léxicas. Dichas probabilidades léxicas junto con el modelo de lenguaje ECGI extendido constituyen la entrada para el etiquetador. Finalmente, para cada frase de entrada, se buscará la secuencia de estados de mayor probabilidad en el modelo ECGI extendido mediante el algoritmo de Viterbi. Esta secuencia óptima tendrá una única categoría léxica que es la etiqueta léxica a devolver como salida. Por último se añade la información morfológica que ha sido extraída en la entrada y se reescribe la oración añadiendo a cada palabra su lema y la etiqueta léxico-morfológica PAROLE ya desambiguada.

El analizador léxico-morfológico toma como entrada un texto plano y obtiene como salida la relación de palabras con sus lemas y una etiqueta que indica la categoría gramatical de la palabra junto con su información morfológica. El conjunto de etiquetas léxico morfológicas contiene unas 230 etiquetas estructuradas en categoría y subcategoría gramatical, y contempla aspectos morfológicos de género, número, persona y tiempos verbales.

3.3 WordNet

La base de conocimiento léxica utilizada a lo largo de esta Tesis ha sido la versión 1.6 de WordNet. Este ha sido el recurso externo utilizado en el método de desambiguación léxica propuesto en

este trabajo y se dispone gratuitamente en formato electrónico en Internet[5].

El propósito fundamental de una base de datos léxica es almacenar información relativa a un conjunto de términos de una o más lenguas. En los últimos años se han desarrollado una serie de proyectos centrados en la construcción de grandes recursos de uso general con información relativa al léxico completo de uno o varios idiomas. Como ejemplo de estos tipos de proyectos se pueden citar a WordNet (Miller et al., 1990; Fellbaum, 1998) y a EuroWordNet (Vossen, 1998). El objetivo del proyecto WordNet es construir un diccionario que permita búsquedas conceptuales en lugar de alfabéticas, inspirándose su diseño en teorías psicolingüísticas sobre la memoria léxica humana. WordNet es una base de datos léxica con información sobre palabras pertenecientes a cuatro categorías sintácticas: nombres, verbos, adjetivos y adverbios. Por su parte, el proyecto EuroWordNet[6] se centra en la construcción de una base de datos con relaciones semánticas entre palabras de varias lenguas Europeas (alemán, español, francés, italiano, estonio y checo). Es decir, EuroWordNet es una base de datos léxica multilingüe, que tiene las palabras de los distintos idiomas enlazadas con la versión 1.5 de WordNet (base de datos léxica en inglés). Mientras que las palabras de los distintos idiomas están mantenidas en base de datos individuales.

Al elemento básico utilizado por WordNet para representar conceptos como conjuntos de sinónimos se le denomina *synset*. En ellos se almacena la información relativa a las diferentes relaciones definidas entre palabras y otros conceptos. Es decir, cada *synset* es una lista de palabras sinónimas y relaciones a otros *synsets*. Estas relaciones pueden ser de dos tipos: léxicas y semánticas. Las relaciones léxicas tratan la forma de las palabras y las relaciones semánticas tratan el significado de las palabras. Entre las relaciones semánticas para nombres se encuentran las de sinonimia, hiperonimia, hiponimia, meronimia, holonimia y términos coordinados. A continuación se presentarán solamente las relaciones de sinonimia, hiperonimia y hiponimia para nombres ya que son las

[5] `http://www.cogsci.princenton.edu/~` *wn*

[6] `http://www.hum.uva.nl/~` *ewn/*

que se utilizan en el método de desambiguación propuesto en esta Tesis.

- **Sinonimia**. La relación de sinonimia es la base para definir el objeto básico de WordNet denominado *synset*. Por lo tanto WordNet une en un mismo *synset* a aquellas palabras que tienen un significado común (sinónimas), formando un item que puede corresponderse intuitivamente con un concepto. Una palabra puede formar parte de varios *synsets*, y esa palabra en cada uno de los diferentes *synsets* representaría un significado distinto. Como muestra el ejemplo 3, la palabra *tree* se encuentra asociada a dos *synsets* o conceptos, los cuales indican sus significados en inglés. La información mostrada en el ejemplo 3 corresponde al formato que devuelve la base de datos léxica WordNet cuando devuelve los sentidos de una determinada palabra.

(3) {tree#1} – (a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown; includes both gymnosperms and angiosperms).

{tree#2, tree diagram#1} – (a figure that branches from a single root; "genealogical tree")

Como se puede observar, esta información se divide en el *synset* y en la "glosa". El *synset* corresponde a la información encerrada entre llaves, es decir {*tree#2, tree diagram#1*}, la cual indica que la palabra *tree* con sentido 2 tiene un significado común a la palabra *tree diagram* con sentido 1. La otra información presentada es la "glosa", la cual define el significado de la palabra igual que lo hace un diccionario. Para indicar cada sentido de la palabra WordNet utiliza el símbolo "#" (número) y a continuación el sentido asociado a la palabra
El WordNet español tiene el mismo formato de salida que el de inglés pero los *synsets* están escritos en español. Como se puede ver en el ejemplo 4, la palabra "árbol" tiene tres sentidos y la glosa de cada uno de ellos está descrita en inglés.

(4) {árbol#1} – (a figure that branches from a single root; "genealogical tree")

{árbol#2, cigüeñal#1} – (a rotating shaft driven by (or driving) a crank)

{árbol#3} – (a tall perennial woody plant having a main trunk and branches forming a distinct elevated crown; includes both gymnosperms and angiosperms)

- **Hiperonimia.** Es el término específico utilizado para designar a un concepto completo de instancias específicas. Un concepto representado por el *synset* x es un hiperónimo del concepto representado por y, si el concepto y es (un tipo de) x. Por ejemplo, $\{woody\ plant\#1\}$ es un hiperónimo de $\{tree\#1\}$ porque un "árbol" es un tipo de "planta de tallo duro". El ejemplo 5 muestra la relación de hiperonimia para $\{tree\#1\}$.

(5) =>{tree#1}
=> {woody plant#1, ligneous plant#1}
=> {vascular plant#1, tracheophyte#1}
=> {plant#2, flora#2, plant life#1}
=> {life form#1, organism#1, being#2}
=> {entity#1, something#1}

- **Hiponimia.** Es definida como la relación inversa de hiperonimia. Es el término específico utilizado para designar a un miembro de un concepto. Un concepto representado por el *synset* x es un hipónimo del concepto representado por y, si el concepto x es (un tipo de) y. Por ejemplo, $\{bonsai\}$ es un hipónimo de $\{tree\}$ porque un *bonsai* es un tipo de "árbol". El ejemplo 6 muestra la relación de hiponimia para $\{tree\#1\}$.

(6) => {tree#1}
=> {yellowwood#3, yellowwood tree#1}
=> {lancewood#2, lancewood tree#2}
=> {Guinea pepper#2, negro pepper#1}
=> .....
=> .....
=> {gymnospermous tree#1}
=> {angiospermous tree#1, flowering tree#1}
=> {fever tree#1}
=> {bonsai#1}

3.4 Método de desambiguación léxica

El método de desambiguación léxica propuesto utiliza como entrada la salida del analizador léxico-morfológico (*TreeTagger*) donde cada palabra contiene su categoría gramatical e información morfológica. Así, las oraciones de entrada previamente etiquetadas se filtran a través de un programa que obtiene únicamente todas aquellas palabras cuya categoría gramatical es nombre. Por lo tanto, la entrada al método será la lista de nombres que forman la oración.

A continuación se explicará detalladamente el método completo para resolver la ambigüedad léxica.

El método completo propuesto para resolver la ambigüedad léxica se compone de dos partes, como se puede ver en la figura 3.1. La primera parte se denomina Método de Marcas de Especificidad, y se aplica al conjunto inicial de palabras que se quieren desambiguar para resolver su ambigüedad léxica. La segunda parte se denomina Conjunto de Heurísticas, y se aplican al conjunto de palabras que no han podido ser desambiguadas por la primera parte con el objetivo de mejorar su desambiguación.

3.4.1 Método de Marcas de Especificidad

Descripción intuitiva. El método de Marcas de Especificidad consiste en desambiguar automáticamente el sentido de las palabras que aparecen dentro del contexto de una oración (microcontexto ó contexto local) mediante la utilización de la taxonomía de nombres de la base de conocimiento léxico WordNet. Este método está basado en la hipótesis de que las palabras que aparecen en un mismo contexto tienen sus sentidos relacionados entre sí. Por lo tanto, se deduce que el contexto juega un papel muy importante a la hora de identificar el significado de una palabra polisémica. Así, en el método propuesto de Marcas de Especificidad el contexto es observado como el grupo de palabras que se encuentran en una ventana circundante (oración) a la ocurrencia de la palabra a desambiguar en la oración, sin considerar la aplicación a dichas palabras de las técnicas de distancia entre términos,

preferencias de selección, relaciones gramaticales, colocación de los sintagmas, etc. En el ejemplo 7 se muestra la oración de entrada (O) y el conjunto de palabras consideradas como contexto (C).

(7) **O:** Baseball, competitive game of skill played with a hard ball and bat between two teams of players each.

C: {baseball, game, skill, ball, team, player}.

El método de Marcas de Especificidad para resolver la ambigüedad léxica coteja el contexto de la palabra a desambiguar con información de un recurso de conocimiento externo. Por lo comentado anteriormente y con motivo de resolver el problema, interesa tener un recurso que tenga las palabras y los conceptos organizados alrededor de clases (jerarquías), de tal forma que describan todas sus características semánticas. El método propuesto fue diseñado para que obtuviera las ventajas comentadas en la sección anterior, por eso se usarán las relaciones jerárquicas de hiperonimia y hiponimia que proporciona la base de conocimiento léxica denominada WordNet. En la Tabla 3.3 se muestran las palabras que forman el contexto de la oración del ejemplo 7 junto a los conceptos que proporciona el recurso externo WordNet.

PALABRA	Concepto WordNet
baseball	{baseball, baseball game, ball}
game	{game}
skill	{skill, accomplishment, acquirement}
ball	{ball}
team	{team, squad}
player	{player, participant}

Tabla 3.3. Palabras del contexto junto a conceptos de WordNet

A continuación se explicará intuitivamente la noción de Marca de Especificidad y como se aplica para realizar la desambiguación del sentido de las palabras.

El método de Marcas de Especificidad está basado en la hipótesis anteriormente comentada: "las palabras en un mismo contexto están fuertemente relacionadas" de este modo si dos o más palabras pertenecen a una misma clase (entendiendo como clase semántica la organización en conceptos en una taxonomía proporcionada) quiere decir que sus sentidos están fuertemente relacionados. Sobre la base anterior, y partiendo de la jerarquía semántica de WordNet, se puede observar que la palabra *baseball y ball* del ejemplo 7 estarían fuertemente relacionadas. Es decir, WordNet a estas dos palabras las representa en su jerarquía semántica mediante una relación de *hiperonimia/hiponimia*, lo que indica que sus sentidos están muy relacionados. De ello se extrae que, cuanta más información común compartan dos conceptos, más relacionados estarán, y la información común que comparten esos dos conceptos se indicará a través de la relación de hiperonimia/hiponimia de ambos en la jerarquía, al cual llamaremos Marca de Especificidad (ME).

Por ejemplo, las palabras *baseball_player y player* comparten la siguiente información común: una persona que participa en competiciones deportivas y en este caso la competición sería *baseball.* Por lo tanto, estas dos palabras están fuertemente relacionadas a través de la jerarquía de WordNet mostrada en la figura 3.3. Si se observa dicha jerarquía, el concepto {*baseball player#1, ballplayer#1*} es un tipo determinado de {*player#1, participant#2*} por lo tanto comparten información común que se puede representar mediante una marca que especifique el significado de cada uno. En este caso la marca de especificidad será el concepto {*player#1, participant#2*} ya que es el concepto superior en la jerarquía.

El modo de proceder del método de Marcas de Especificidad consistirá en recorrer todos los subárboles de la jerarquía semántica de WordNet para cada una de las palabras que forman el contexto de entrada y para cada una de las Marcas de Especificidad calculará cuantas palabras del contexto de entrada se agrupan alrededor de ella. Aquella Marca de Especificidad que agrupe el máximo número de palabras del contexto, será elegida como el sentido de la palabra. Dicho de otro modo, el método busca aquella Marca de Especificidad que tenga mayor densidad

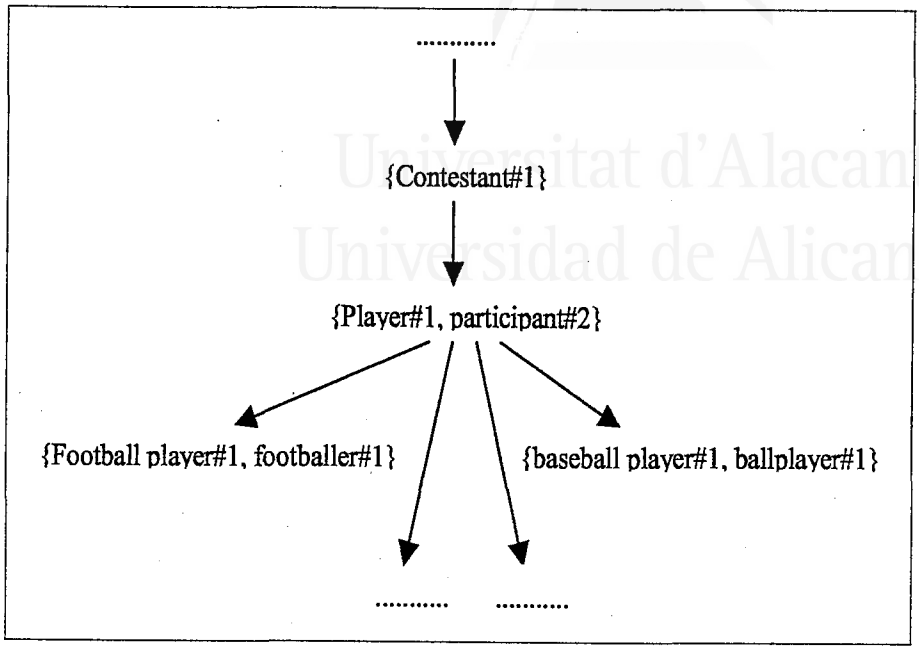

Figura 3.3. Jerarquía para player y baseball

de palabras debajo de su subárbol, lo cual quiere decir que sus sentidos están fuertemente relacionados. De este modo el proceso se aplicará sobre la jerarquía de WordNet del siguiente modo.

La entrada al método de WSD estará formada por el conjunto de palabras $W = \{w_1, w_2, ..., w_n\}$ que se obtienen de la oración y forman el contexto. Cada palabra w_i se busca en WordNet y se obtienen los sentidos asociados a ellas $S_i = \{s_{i1}, s_{i2}, ..., s_{in}\}$ y para cada sentido s_{ij} se obtendrá todos los conceptos (*synset*) hiperonimos en su taxonomía IS-A. Inicialmente, se busca el concepto común a todos los sentidos de las palabras que forman el contexto de entrada, denominándose a este concepto Marca de Especificidad Inicial (MEI). Si esta Marca de Especificidad Inicial no resuelve la ambigüedad de las palabras, se va descendiendo nivel a nivel a través de la jerarquía WordNet asignando nuevas Marcas de Especificidad. Para cada Marca de Especificidad anterior, se calculará el número de conceptos que forman parte del contexto y que están contenidos en la subjerarquía. Aquella Marca de Especificidad que en su subjerarquía, tenga el mayor número de

palabras del contexto será la elegida, asignando el sentido que nos devuelve WordNet a cada una de estas palabras que forman parte de la Marca de Especificidad seleccionada.

En la figura 3.4 se muestra un ejemplo que ilustra de forma intuitiva y gráfica como trabaja el método de WSD mediante Marcas de Especificidad. En el gráfico no se han tenido en cuenta los sentidos 3 y 4 de la palabra *plant*.

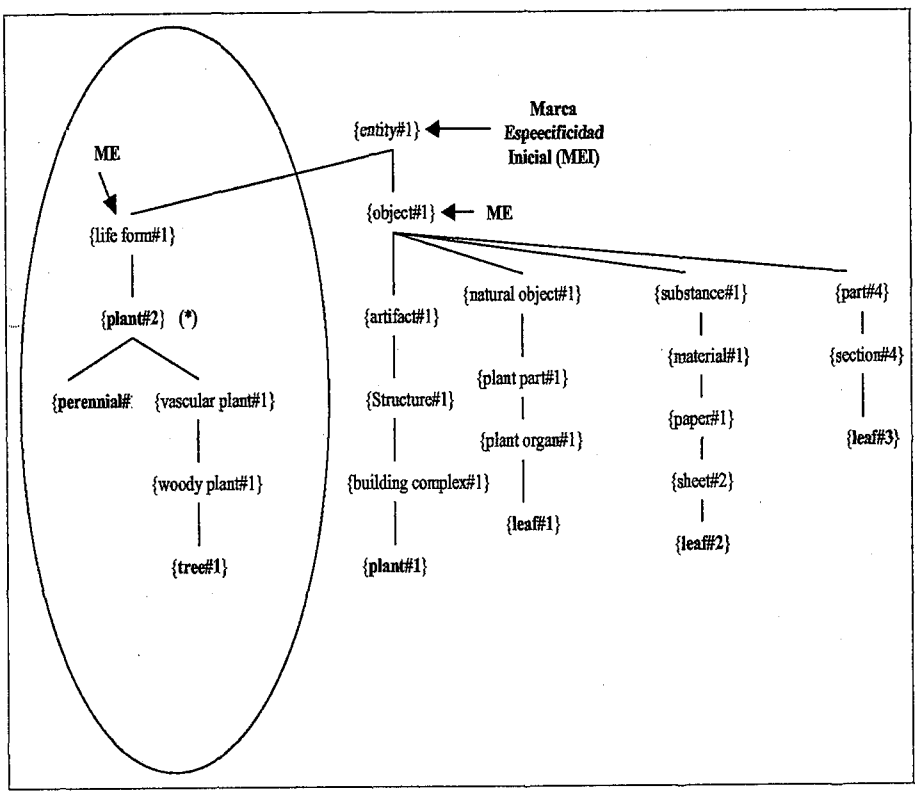

Figura 3.4. Método WSD usando Marcas de Especificidad

Como se puede apreciar el conjunto de palabras (en nuestro caso nombres) que forman el contexto de entrada a partir de una oración es el siguiente: {*plant, tree, perennial, leaf*}. Según WordNet 1.6, la palabra *plant* tiene cuatro sentidos diferentes, *tree* dos, *perennial* uno y *leaf* tres. Por lo tanto, si se pretende desambiguar *plant*, el método tratará a las otras tres palabras como contex-

to, y esto mismo se aplicará al resto de palabras cuando vayan a desambiguarse. La Marca de Especificidad Inicial de la figura no resuelve la ambigüedad léxica, ya que las palabras *plant, tree y leaf* aparecen en cuatro, dos y tres subjerarquías con diferentes sentidos, respectivamente. Por tal motivo el método debe ir descendiendo nivel a nivel a través de la jerarquía mostrada, asignando nuevas Marcas de Especificicidad. Sin embargo, la Marca de Especificidad con el símbolo (*) contiene el mayor número de palabras del contexto (tres) en su subjerarquía y, por lo tanto, será elegida para resolver con el sentido {*plant#2, flora#2, plant life#1*} la palabra *plant*. Las palabras *tree* y *perennial* también son desambiguadas eligiendo el sentido {*tree#1*} y {*perennial#1*} para ambas. Sin embargo, para la palabra *leaf*, no ha sido desambiguada satisfactoriamente porque WordNet no la relaciona en su taxonomía con el resto de palabras. Cuando se da este caso, entonces se aplicarán un conjunto de heurísticas auxiliares. Esta heurísticas se presentarán en la siguiente sección.

A continuación, una vez presentado de forma intuitiva el proceso del método de Marcas de Especificidad, se describirán detalladamente cada uno de las 5 fases que se ha dividido dicho método.

- FASE 1: **Entrada.**
  El método toma como entrada conjuntos de palabras a desambiguar correspondiente a cada una de las oraciones del documento. En concreto el conjunto de palabras serán el conjunto de nombres de cada oración; de este modo cada conjunto formará el denominado contexto. El ejemplo 7 muestra una posible entrada al método.
- FASE 2: **Obtención de sentidos e hiperónimos.**
  Para cada una de las palabras anteriores pertenecientes a un contexto, se obtienen todos sus sentidos que suministra la base de datos WordNet y para cada sentido se obtienen todos los conceptos hiperónimos suministrados por la misma base de datos. Un ejemplo de este paso sería el mostrado en la Tabla 3.4. Se puede observar que la palabra *plant* tiene 4 sentidos distintos

en WordNet 1.6 y cada uno de esos sentidos tienen asociados sus conceptos hiperónimos.

plant#1	**plant#2**	**plant#3**	**plant#4**
plant#1	plant#2	plant#3	plant#4
building complex#1	life form#1	contrivance#3	actor#1
structure#1	entity#1	scheme#1	performer#1
artifact#1		plan of action#1	entertainer#1
object#1		plan#1	person#1
entity#1		idea#1	life form#1
		content#5	entity#1
		cognition#1	
		psychological feature#1	

Tabla 3.4. Hiperónimos de los sentidos de *Plant*

- FASE 3: **Recorrido de la red de hiperonimia.**
  El objetivo de esta fase es construir una estructura de datos para cada uno de los sentidos de las palabras que se quieren desambiguar. En la Figura 3.5 se muestra un ejemplo que describe la información obtenida al aplicar este paso para los sentidos 1 y 2 de la palabra *plant*. Para la construcción de esta estructura, hay que recorrer la red de hiperonimia de WordNet con el objetivo de encontrar si cada uno de los conceptos hiperónimos de los sentidos de las palabras a desambiguar, obtenidos en la fase anterior, contiene o es clase superior de los conceptos hiperónimos correspondientes al resto de sentidos del conjunto de palabras que forman el contexto. Posteriormente, se almacenará para cada uno de los sentidos de las palabras a desambiguar una estructura de datos en forma de lista con toda la información obtenida en la búsqueda de cada concepto hiperónimo. Hay que aclarar que esto se debería hacer para todas las palabras que forman el contexto y para cada uno de los sentidos de esas palabras. La parte izquierda de la lista son las distintas Marcas de Especificidad del método.

Para PLANT:
  Para PLANT#1:
    plant#1 →
    building complex#1 →
    structure#1 →
    artifact#1 →
    object#1 → leaf#1, leaf#2, leaf#3
    entity#1 → plant#2, plant#4, tree#1, perennial#1, leaf#1, leaf#2, leaf#3
  Para PLANT#2:
    plant#2 → tree#1, perennial#1
    life form#1 → tree#1, perennial#1, plant#4
    entity#1 → plant#1, plant#4, tree#1, perennial#1, leaf#1, leaf#2, leaf#3

Figura 3.5. Estructura de datos para dos sentidos de *plant*

- FASE 4: **Obtención de la densidad de palabras.**
  El objetivo de esta fase es obtener la densidad de palabras para cada uno de los sentidos de las palabras a desambiguar y para cada uno de los niveles de la jerarquía de WordNet. Para obtener la densidad de las palabras, hay que contar el número de palabras que se relacionan con cada uno de los sentidos de las palabras que se quieren desambiguar, a partir de la estructura de datos en forma de lista obtenida en el paso anterior. Para cada sentido hay que posicionarse en la Marca de Especificidad del nivel[7] superior y contar el número de palabras que contiene. En la figura 3.6 se muestra un ejemplo que describe el número de palabras que forman parte de cada uno de los sentidos según el contexto dado, estando posicionados en la Marca de Especificidad del nivel superior, posteriormente se pasará al siguiente nivel y así sucesivamente hasta llegar al nivel inferior. El resultado de aplicar lo comentado anteriormente a la palabra *plant* muestra que los sentidos 1, 2 y 4 tienen la misma densidad de palabras para la Marca de Especificidad del último nivel. Por lo tanto, esa palabra no se puede desambiguar estando en esa Marca de Especificidad con lo que habrá que cambiar de Marca

[7] El último nivel en la lista, será el nivel de menor profundidad en la jerarquía de hiperónimos (entity, abstract, etc)

de Especificidad subiendo un nivel y seguir el recorrido de la red de hiperonimia.

Para PLANT
Para PLANT#1 : 4 (plant, tree, perennial, leaf)
Para PLANT#2 : 4 (plant, tree, perennial, leaf)
Para PLANT#3 : 1 (plant)
Para PLANT#4 : 4 (plant, tree, perennial, leaf)

Figura 3.6. Número de palabras para los sentidos de *plant*

- FASE 5: **Selección de sentidos (salida).**
  El objetivo de esta fase es elegir aquel sentido de la palabra que tenga el mayor valor de densidad de palabras obtenidos en la fase 4 (obtención densidad de palabras). Pero cuando se realiza esta acción pueden ocurrir tres casos diferenciados:
  - Un único sentido de la palabra con el mayor valor, entonces se elige ese sentido.
  - Más de un sentido de la palabra con el mismo valor máximo, entonces se volverá a aplicar la fase 4 (obtención densidad de palabras), con el cambio de Marca de Especificidad a un nivel inferior en la estructura de datos de la fase 3 (recorrido de la red de hiperonimia) hasta obtener un único sentido. Un ejemplo de este caso sería el mostrado en la figura 3.6, donde los sentidos 1, 2 y 4 para *Plant* tiene el mismo valor (cuatro) como valor máximo posicionados en la Marca de Especificidad de *entity*.
  - No se pueda obtener un único sentido al aplicar los puntos anteriores, entonces se le asigna una etiqueta denominada de "ambigua", que significa que la palabra no ha podido desambiguarse con las relaciones hiperónicas que contiene WordNet.

  Si se aplica lo comentado anteriormente a los datos obtenidos en la Figura 3.6 se obtienen tres sentidos {*plant#1, plant#2, plant#4*} con el mismo valor máximo de densidad de palabras (4), por lo que no se puede desambiguar. A continuación, observando los datos de la Figura 3.5, habría que cambiar a la Marca

de Especificidad de {*object#1*} y {*life form#1*} para el sentido de {*plant#1*} y {*plant#2*} respectivamente y volver a aplicar la fase 4 (obtención densidad de palabras). Una vez aplicado el mismo procedimiento anterior, en el caso de la Marca de Especificidad {*life form#1*} se obtendría un valor de densidad de 3 palabras, el cual obtiene la máxima densidad para la palabra *plant*, por lo que se elegiría a {*plant#2*} como su sentido correcto.
Un ejemplo del caso que se le asigna a la palabra la etiqueta "ambigua" es cuando la palabra *leaf* forma parte con la palabra *plant* en el contexto de entrada. Y esto es debido a que la base de datos WordNet no relaciona, utilizando la hiperonimia o la hiponimia, la palabra *leaf* con el sentido de "hoja" con la palabra *plant* con el sentido de "flora". Para tratar estos casos se utiliza un conjunto de heurísticas que se detallan en secciones posteriores.

Un ejemplo completo del método de Marcas de Especificidad se muestra a continuación, con el detalle de cada uno de los pasos aplicados.

- **Oración:**

El usuario quiere desambiguar el sentido de los nombres que intervienen en esta oración:

"The bristles are a Fitch 2 and one-half inch brush shaved to a sharp chisel edge".

continua →

- FASE 1: **Entrada.**

  El método toma como entrada el conjunto de nombres obtenidos a partir de la oración anterior. A este conjunto de nombres se le denomina el contexto de la oración. Para nuestro caso serán:

  {bristle, one-half, inch, brush, chisel, edge}.

- FASE 2: **Obtención de sentidos e hiperónimos.**

  Para cada una de las palabras obtenidas en el paso 1 se obtienen todos sus sentidos $S_i = \{s_{i1}, s_{i2}, ..., s_{in}\}$ y para cada sentido s_{ij} se obtienen todos los synsets hiperónimos, almacenándolos en una estructura de datos en forma de pila. (Únicamente se indican los sentidos de la palabras bristle).

  $S_1 = \{bristle\#1, bristle\#2\}$

  A continuación, se presentan los hiperónimos de la palabra Bristle para cada uno de sus sentidos:

Estructura Hiperónimos (Bristle)	
bristle#1	bristle#2
fiber#1	hair#2
material#1	process#6
sunstance_matter#1	body_part#1
object#1	part#7
entity#1	entity#1

  Para cada uno de los sentidos de las siguientes palabras se debería hacer lo mismo.

  $S_2 = \{one\_half\#1\}$
  $S_3 = \{inch\#1, inch\#2\}$
  $S_4 = \{brush\#1, brush\#2, \ldots, brush\#8\}$
  $S_5 = \{chisel\#1\}$
  $S_6 = \{edge\#1, edge\#2, \ldots, edge\#6\}$

continua →

- FASE 3: **Recorrido de la red de hiperonimia.**

Para BRISTLE

Para BRISTLE#1:
bristle#1 →
fiber#1 →
material#1 →
substance_matter#1 →
object → *chisel*#2, *edge*#1, *edge*#3, *edge*#4, *brush*#2, *brush*#4
entity → *chisel*#2, *edge*#1, *edge*#3, *edge*#4, *brush*#2, *brush*#4, *bristle*#2

Para BRISTLE#2:
bristle#2 →
hair#2 →
process#6 →
body_part#1 →
part#7 →
entity#1 → *chisel*#2, *edge*#1, *edge*#3, *edge*#4, *brush*#2, *brush*#4, *bristle*#2

Para CHISEL:

Para CHISEL#1:
chisel#1 →
hand_tool#1 →
tool#1 →
implement#1 → *brush*#2
instrumentation#1 → *brush*#2, *brush*#4
artifact#1 → *brush*#2, *brush*#4, *edge*#3, *edge*#6
object#1 → *brush*#2, *brush*#4, *edge*#3, *edge*#6, *edge*#1, *bristle*#1
entity#1 → *brush*#2, *brush*#4, *edge*#3, *edge*#6, *edge*#1, *bristle*#1, *bristle*#2

Y así para el resto de las palabras.

- FASE 4: **Obtención densidad de palabras.**

Situándonos inicialmente en el último nivel, es decir en $< entity >$, $< abstraction >$, $< group >$, $< event >$, $< act >$ el valor para cada uno de los sentidos de cada una de las palabras a ese nivel es la siguiente:

Para BRISTLE

Para bristle#1 : 4 (chisel, edge, brush, bristle)
Para bristle#2 : 4 (chisel, edge, brush, bristle)

continua →

Para BRUSH

Para brush#1 : 1
Para brush#2 : 4
Para brush#3 : 3
Para brush#4 : 4
Para brush#5 : 1
Para brush#6 : 1
Para brush#7 : 1
Para brush#8 : 1

Para CHISEL
Para chisel#1 : 4

Para ONE_HALF
Para one_half#1 : 3

Para EDGE
Para edge#1 : 4
Para edge#2 : 3
Para edge#3 : 4
Para edge#4 : 3
Para edge#5 : 3
Para edge#6 : 4

Para INCH
Para inch#1 : 3
Para inch#2 : 3

- FASE 5: **Selección de sentidos (salida).**

  Ahora hay que escoger aquellas palabras que tengan el valor máximo y si no se consigue desambiguar la palabra se volverá al paso 4 bajando de nivel.

  Para BRISTLE
  No se desambigua, ya que el valor de ambos es máximo por lo que continúan los dos sentidos bristle#1, bristle#2.

  Para BRUSH
  No se desambigua, ya que el valor de ambos es máximo en los sentidos brush#2 y brush#4, y el resto de sentidos se descartan.

  Para CHISEL
  Si se desambigua ya que únicamente tiene 1 sentido y por lo tanto ese es el resultado.

continua →

Para ONE_HALF
Al igual que chisel, se desambigua al tener un único sentido.

Para EDGE
No se desambigua, los sentidos edge#1, edge#3 y edge#6 poseen el valor máximo, el resto se descartan.

Para INCH
No se desambigua, los dos sentidos que tiene, tienen la cuenta máxima, inch#1, inch#2.

Como aun quedan palabras sin desambiguar, volvemos al paso 4 pero bajando 1 nivel.

- FASE 4: **Obtención densidad de palabras.**

  Bajamos un nivel por lo que nos situamos en: $< object >$, $< part >$, $< measure >$, $< attribute >$, y calculamos de nuevo la densidad para los sentidos no descartados.

  Para BRISTLE
  Para bristle#1 : 4 (nivel de $< object >$)
  Para bristle#2 : 1 (nivel de $< part >$)

  Para BRUSH
  Para brush#2 : 4 (nivel de $< object >$)
  Para brush#4 : 4 (nivel de $< object >$)

  Para EDGE
  Para edge#1 : 1 (nivel de $< object >$)
  Para edge#3 : 4 (nivel de $< object >$)
  Para edge#6 : 4 (nivel de $< object >$)

  Para INCH
  Para inch#1 : 2 (nivel de $< mesure >$)
  Para inch#2 : 2 (nivel de $< mesure >$)

- FASE 5: **Selección de sentidos (salida).**

  Para BRISTLE
  Se desambigua, el resultado es el sentido uno, bristle#1.

  Para BRUSH
  No se desambigua, el valor sigue siendo máximo en los sentidos brush#2 y brush#4.

continua →

Para EDGE
No se desambigua, los sentidos edge#3 y edge#6 poseen el valor máximo pero el sentido uno se elimina.

Para INCH
No se desambigua, el valor sigue siendo máximo para inch#1 e inch#2.

Como aun quedan palabras sin desambiguar y niveles para poder descender, se vuelve al paso 4.

- FASE 4: **Obtención densidad de palabras.**

Para BRUSH
Para brush#2 : 3 (nivel de $< artifact >$)
Para brush#4 : 3 (nivel de $< artifact >$)

Para EDGE
Para edge#3 : 3 (nivel de $< artifact >$)
Para edge#6 : 3 (nivel de $< artifact >$)

Para INCH
Para inch#1 : 1 (nivel de $< linear\_measure >$)
Para inch#2 : 2 (nivel de $< definite\_quantity >$)

- FASE 5: **Selección de sentidos (salida).**

Para BRUSH, no se desambigua.

Para EDGE, no se desambigua.

Para INCH, se desambigua con inch#2.

El algoritmo continuaría aplicando las fases comentadas (tres, cuatro y cinco) hasta llegar al *synset* hiperónimo $< implement >$ y paralelamente $< device >$ donde se consigue desambiguar *brush* dando como resultado el sentido *brush#2*. El nombre *edge* no se consigue desambiguar ya que llega un momento que para todos los hiperónimos de ambos sentidos el valor es 1 por lo que se deben aplicar las heurísticas, que se detallarán a continuación, para conseguir su desambiguación completa.

Los resultados dados por el algoritmo para cada una de las palabras son los siguientes:

Palabras	Sentido ME	Sentido SemCor
bristle	# 1	# 1
one half	# 1	# 1
inch	# 2	# 1
brush	# 2	# 2
chisel	# 1	# 1
edge	# 3	# 3

Tabla 3.5. Resultados de desambiguación

Como se observa en la tabla de arriba, todas las palabras han sido desambiguadas con sus correspondientes sentidos y solamente el sentido asignado a la palabra *inch* es incorrecto, ya que Semcor le asigna el sentido 1 y nuestro método le asigna el sentido 2. Esto ha pasado porque los dos sentidos asignados a esta palabra en WordNet tienen una jerarquía semántica tan fina que el método no ha podido desambiguar correctamente.

Descripción formal. Una vez descrito intuitivamente el método, a continuación se describirá formalmente el algoritmo de Marcas de Especificidad.

Dada una oración O, el algoritmo extrae todas aquellas palabras cuya categoría léxica sea nombre. Estos nombres constituyen el contexto de entrada $C = \{w_1, w_2, \ldots, w_n\}$. Para cada una de estas palabras, se hace una llamada a la base de conocimiento WordNet para extraer toda la información de sus sentidos y la estructura de hiperónimos que cada uno contenga. WordNet devuelve toda esta información en una estructura de datos de lista enlazada como se muestra en la figura 3.7.

Para obtener la información a partir de esta estructura se utilizan dos punteros. El puntero denominado *nextss*, que avanza por todos los sentidos de la palabra, y el puntero *ptrlist* que avanza por los hiperónimos. En la descripción formal del algoritmo que se detallará a continuación, a la estructura mostrada se le denominará *Lista_completa*.

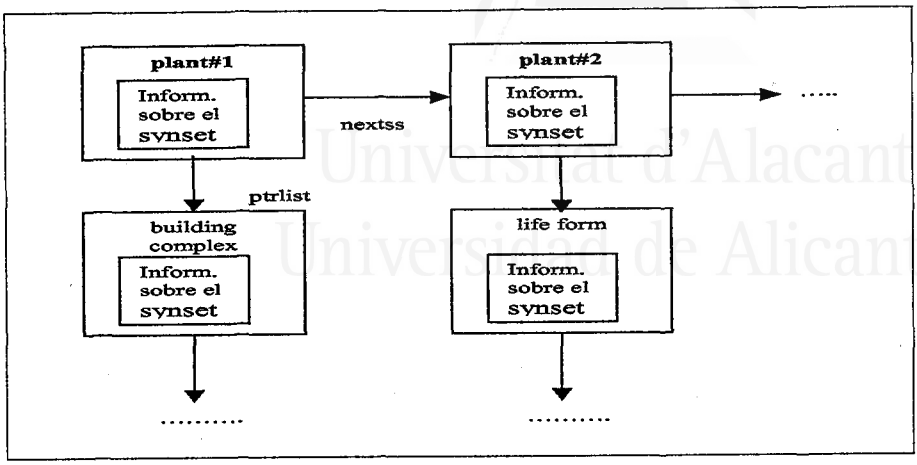

Figura 3.7. Estructura de datos en forma de lista enlazada para la palabra *plant*

Una vez comentadas las estructuras de datos internas que tiene WordNet, a continuación se describirán las distintas funciones utilizadas en el algoritmo. En la Figura 3.8 se presenta el algoritmo principal con las llamadas a cada una de las funciones y sus parámetros de entrada y de salida.

Algoritmo Marcas_Especificidad

---

obtener_sentidos(Contexto(C), Lista_enlazada);
completar_lista(Lista_enlazada, Lista_completa);
calcular_numero_palabras(Lista_completa, Valores_lista);
elegir_sentido(Valores_lista, Sentidos)

Fin Algoritmo

Figura 3.8. Algoritmo principal

Una vez detalladas las funciones que se aplican en el algoritmo, a continuación se describirán formalmente cada una de las mismas en las Figuras 3.9, 3.10, 3.11 y 3.12. La información que se detalla

para cada función es: el nombre de la función, los parámetros de entrada, los de salida y una descripción formal de la actividad que desempeñan.

Función obtener_sentidos

Entrada: Conjunto de nombres que forman el contexto $C = \{w_1, w_2, \ldots, w_n\}$

Salida: Lista_enlazada

Descripción: Esta función obtiene la información referente a cada sentido en WordNet con todos sus hiperónimos.

Algoritmo:

Para cada palabra w_i
  Obtener todos sus sentidos s_{ij} de WordNet
  Para cada sentido s_{ij}
    Almacenar Hiperonimo en Lista_enlazada
  Fin para
Fin para

Figura 3.9. Función para obtener los distintos sentidos de las palabras

3.4.2 Heurísticas.

Evidentemente, las relaciones de hiperonimia y de hiponimia no capturan todas las relaciones semánticas que ocurren entre las palabras de un texto. Por eso, al aplicar el método de Marcas de Especificidad a diversos ejemplos se observó que había palabras en el contexto de entrada que estaban muy relacionadas entre sí semánticamente (por ejemplo *plant* y *leaf*). Por tal motivo, y

Función completar_lista

Entrada: Lista_enlazada

Salida: Lista_enlazada_completa

Descripción: Esta función busca para cada sentido sus hiperónimos en todos los demás sentidos del resto de palabras, completando la estructura inicial de Lista_enlazada.

Algoritmo:

Para cada palabra w_i
  Para cada sentido s_{ij}
    Para cada $Hiperonimo_{ij}$ de s_{ij}
      Buscar este $Hiperonimo_{ij}$ en todos los demás sentidos del resto de palabras del conjunto C.
      Si aparece este $Hiperonimo_{ij}$ **entonces**
        Almacenar sentidos de las palabras que tienen ese hiperónimo en Lista_enlazada_completa
      Fin si
    Fin para
  Fin para
Fin para

Figura 3.10. Función para construir la estructura dato lista

después de realizar diversos estudios sobre el método se observó que los resultados de desambiguación se podían mejorar definiendo una serie de heurísticas con el objetivo de ayudar a resolver el resto de relaciones semánticas. A cada una de las heurísticas definidas para mejorar el método se les ha asignado un nombre representativo de la acción que realizan, siendo las heurísticas del Hiperónimo, de Definición, de Hipónimo, de Glosa Hiperónimo,

Función calcular_numero_palabras

Entrada: Lista_enlazada_completa
Salida: Lista_enlazada_completa con valores

Descripción:
Esta función calcula la densidad de palabras para cada uno de los sentidos de las palabras a desambiguar y para cada uno de los niveles de la jerarquía de WordNet. Para obtener la densidad de las palabras, hay que contar el número de palabras que se relacionan con cada uno de los sentidos de las palabras que se quieren desambiguar, a partir de la estructura de datos Lista_enlazada_completa.

Algoritmo:

Para cada nivel n_h
  Para cada palabra w_i
    Para cada sentido s_{ij}
      Contar en la Lista_enlazada_completa el número de palabras en el nivel n_h
      Asignar número al sentido s_{ij}
    Fin para
  Fin para
Fin para

Figura 3.11. Función para calcular número de palabras en la lista

de Glosa Hipónimo y de Marca Especificidad Común. Siguiendo con la misma filosofía que en el método, para explicar las distintas heurísticas también se realizará una descripción intuitiva y otra formal. Hay que aclarar que estas heurísticas se activan siempre y cuando una vez aplicado el método de Marcas de Especificidad quedan palabras por desambiguar, porque de esta forma se obtienen mejores resultados.

Función elegir_sentido

Entrada: Lista_enlazada_completa con valores
Salida: sentido a cada palabra

Descripción: Esta función selecciona aquel sentido cuyo valor calculado anteriormente sea máximo para cada una de las palabras.

Algoritmo:

Para cada palabra w_i
En caso de
Hay un único sentido **entonces**
Elegir ese Sentido
Hay más de un sentido y uno es máximo **entonces**
Elegir ese Sentido
Hay más de un sentido y valores iguales **entonces**
Subo un nivel en la *Lista_enlazada_ij*
Obtener número de palabras de ese nivel en
Lista_enlazada_completa con valores
En otro caso
Asignar la etiqueta Contexto
Mostrar todos los sentidos
Elegir el más frecuente de WordNet.
Fin en caso
Fin para

Figura 3.12. Función para elegir el sentido correcto

Descripción intuitiva.

- *Heurística del Hiperónimo.* Esta heurística se utiliza para resolver la ambigüedad de aquellas palabras, que formando parte del contexto, no están directamente relacionadas por WordNet (*plant* y *leaf*, etc). Pero, sin embargo, a veces si que aparece alguna de las palabras del contexto como miembro de un *synset* de alguna relación de hiperonimia para algún sentido de la palabra a ser desambiguada. Por ejemplo, si lo que se quiere desambiguar son las palabras {*plant,leaf*} con el método de Marcas de Especificidad se tiene el problema de que estas dos palabras no están directamente relacionadas por WordNet. Pero WordNet si que relaciona la palabra *leaf#1* con *plant_organ#1* a través de la relación de hiperonimia, por lo tanto como el sentido 1 de *leaf* está relacionado con *plant_organ#1* y esta palabra está compuesta de *plant*, una de las palabras que forman el contexto entonces se elige *leaf#1* como sentido correcto desechándose el resto de sentidos.
  Esta heurística actúa del siguiente modo para desambiguar una palabra dada: se cotejan los hiperónimos de sus sentidos buscando las palabras del contexto. Si se encuentra algún *synset* hiperónimo conteniendo palabras del contexto, se asigna un peso en relación a su profundidad de la subjerarquía. El sentido con mayor peso será el elegido como correcto. Un ejemplo de esta heurística se muestra en la Figura 3.13. En este ejemplo se presentan las palabras que forman el contexto, las que no se han podido desambiguar y de estas últimas palabras todos los sentidos posibles (se le llama sentidos finales). Como puede observarse, en los hiperónimos del sentido *leaf#1* aparece *plant* que es otra de las palabras pertenecientes al conjunto de entrada o contexto. El peso para este sentido sería el siguiente:

$$peso = \sum_{i=1}^{profundidad} \left(\frac{N^o\ nivel}{N^o\ niveles\ total}\right) = \left(\frac{4}{6}\right) + \left(\frac{5}{6}\right) = 1,5.$$

Palabras del contexto: *plant, tree, leaf, perennial*
Palabras no desambiguadas: *leaf.*
Sentidos finales: *leaf#1, leaf#2, leaf#3.*

Para leaf#1

=> entity, something	*Nivel 1*
=> object, physical object	*Nivel 2*
=> natural object	*Nivel 3*
=> **plant** part	*Nivel 4*
=> **plant** organ	*Nivel 5*
=> leaf#1, leafage, foliage	*Nivel 6*

Figura 3.13. Aplicación de la Heurística del Hiperónimo

- *Heurística de Definición.* WordNet incluye definiciones (glosas[8]) para cada sentido asociado a una palabra. Estas definiciones son útiles porque tienen asociadas un micro-contexto[9] para cada sentido.
  Esta heurística actúa del siguiente modo para desambiguar una palabra dada: se cotejan las definiciones asociadas a cada sentido que nos suministra WordNet con las palabras que forman el contexto. Cada vez que coinciden las palabras en la definición de un sentido, se incrementa su peso en una unidad. El sentido que finalmente tiene el mayor peso es elegido. Un ejemplo de esta heurística se muestra en la Figura 3.14.
- *Heurística de Marca de Especificidad Común.* Esta heurística trata de resolver el problema de la granularidad fina[10] (*line, have, year, month, etc*).
  Esta heurística actúa del siguiente modo para desambiguar una palabra dada: se elige aquella Marca de Especificidad que sea común en la jerarquía a todos los sentidos ambiguos, ya que

[8] Glosa es la definición correspondiente a cada uno de los *synsets* que forman la base de datos léxica WordNet. Un ejemplo en inglés es: {tree#2, tree diagram#1} – (a figure that branches from a single root; "genealogical tree"). Otro ejemplo, pero para español sería: {ojo#2} – (a small hole or loop (as in a needle).

[9] Se considera micro-contexto al conjunto de palabras que están en una ventana de tamaño variable de texto y que todas están relacionadas con la acción que especifican.

[10] Se dice que una palabra tiene granularidad fina, cuando a esta se le asocian diversos sentidos, que a veces cuesta mucho distinguir cual es el correcto.

Palabras del contexto: *person, sister, musician.*
Palabras no desambiguadas: *sister,musician.*
Sentidos finales: *sister#1, sister#2, sister#3 sister#4.*

Para sister#1 → peso = **2**

1. sister, sis -- (a female **person** who has the same parents as another **person**; "my sister married a **musician**")

Para sister#3 → peso = **1**

3. sister -- (a female **person** who is a fellow member (of a sorority or labor union or other group); "none of her sisters would betray her")

Figura 3.14. Aplicación de la Heurística de definición

proporciona el concepto común más informativo a tales sentidos. Mediante esta heurística se intenta resolver el problema de la granularidad fina que posee WordNet, ya que en la mayoría de los casos, los sentidos resultantes de las palabras se diferencian en un pequeño matiz y debido a que el contexto es muy general no se consigue dar con el sentido más adecuado. Un ejemplo de esta heurística se muestra en la Figura 3.15.
Como puede observarse en el ejemplo de la Figura 3.15, debido a la granularidad tan fina que tiene la versión de WordNet 1.6 y que la palabra *month* no especifica nada sobre alguno de los sentidos de la palabra *year*, lo que más se puede afinar es diciendo que la marca de especificidad común para el sentido de *year* es *time period.*

- *Heurística del Hipónimo.* Heurístico simétrico al del hiperónimo que intenta relacionar las palabras que forman parte del contexto con algún *synset* compuesto (*signal_fire*) de una relación de hiponimia para algún sentido de la palabra a ser desambiguada. Por ejemplo, para la palabra *sign* con sentido 3 al profundizar dos niveles mediante la relación de hiponimia hay un *synset* formado por *watch_ fire.*
Para desambiguar una palabra dada, se busca en las palabras de cada uno de sus *synsets* hipónimos el resto de las palabras

Palabras del contexto: *year, month.*
Palabras no desambiguadas: *year.*
Sentidos finales: *year#1, year#2, year#3.*

Para year#1:

=> abstraction
  => measure, quantity
    => **time period, period**
      => year#1, twelvemonth

Para year#2:

=> abstraction
  => measure, quantity
    => **time period, period**
      => year#2

Figura 3.15. Aplicación de la Heurística de Marca de Especificidad Común

pertenecientes al conjunto de entrada, asignándoles un peso, al igual que hacíamos con la heurística del hiperónimo. El resultado para cada una de las palabras será aquel sentido cuyo peso sea mayor. Un ejemplo de esta heurística se muestra en la Figura 3.16. Como puede observarse en este ejemplo el mejor sentido es *sign#3*, debido a que tiene el mayor peso (1.33).

- *Heurística de la Glosa del Hiperónimo.*
  Otra fuente de conocimiento que proporciona WordNet son sus glosas o definiciones. Esta heurística usa la glosa de cada *synset* de las relaciones de hiperonimia de la palabra a desambiguar que proporciona WordNet.
  Este heurístico busca el resto de las palabras pertenecientes al contexto en la definición de cada *synset* hiperónimo de la palabra a desambiguar. Dependiendo de si existen o no en la subjerarquía a cada uno de los sentidos se les asigna un peso. El resultado para cada una de las palabras será aquel sentido cuyo peso sea máximo. En resumen, esta heurística lo que hace es, buscar en las definiciones de cada *synset* hiperónimo aquellas

Contexto: *ground, fire, sign, activity*
Palabra no desambiguada: *sign*
Sentidos finales: *sign#1, sign#2, sign#3,..., sign#9, sign#10*

Para Sign#2→ Peso=**0**

=> scoreboard
=> poster, placard, notice, bill, card
=> show bill, show card, theatrical poster
=> flash card
=> street sign
=> address
=> signpost, guidepost
=> fingerpost, fingerboard

Para Sign#3→ Peso=(1/2)+(1/2)+(1/3) =**1.33**

=> recording
=> bologram, bolograph
=> chromatogram
=> oscillogram
=> spirogram
...
=> fire alarm
=> foghorn, fogsignal
...

=> visual signal
=> watch **fire**
=> light
...

=> rocket, skyrocket
=> beacon, beacon **fire**
=> signal **fire**, signal light

Para Sign#4→ Peso=**0**

=> billboard, hoarding
=> sandwich board
=> shingle

Figura 3.16. Aplicación de la Heurística del Hipónimo

palabras que forman parte del contexto de entrada. Un ejemplo de esta heurística se muestra en la Figura 3.17. Como puede observarse en este ejemplo el sentido mejor es el *plane#1*, debido a que es el que mayor peso tiene.

- *Heurística de la Glosa del Hipónimo.*
  Esta heurística usa la glosa de cada *synset* de las relaciones de hiponimia de la palabra a desambiguar que proporciona WordNet. Este heurístico busca el resto de las palabras pertenecientes al contexto en la definición de cada *synset* hipónimo de la palabra a desambiguar. Dependiendo de si existen o no en la subjerarquía, a cada uno de los sentidos se les asigna un peso según su profundidad. Así a mayor profundidad menor peso y menor relación con el sentido tratado. El resultado para cada una de las palabras será aquel sentido cuyo peso sea máximo. En resumen, esta heurística lo que hace es, buscar en las definiciones de cada *synset* hipónimo aquellas palabras que forman parte del contexto de entrada. Un ejemplo de esta heurística se muestra en la Figura 3.18. Como puede observarse en este ejemplo el sentido mejor es el *cost#1*, debido a que es el que mayor peso tiene.

Descripción formal. Una vez se ha realizado la explicación intuitiva de cada una de las heurísticas que mejoran al método de Marcas de Especificidad, a continuación se describirán formalmente cada una de ellas mediante las Figuras 3.19, 3.20, 3.21, 3.22, 3.23 y 3.24.

3.4.3 Interfaz Web

En esta sección se presenta el diseño e implementación de una interfaz que resuelve la ambigüedad léxica de nombres mediante la aplicación del método de Marcas de Especificidad descrito en las secciones anteriores. Para realizar la implementación de esta interfaz se eligió la tecnología Internet, ya que así esta interfaz sería accesible desde cualquier punto de la red. La dirección URL para acceder y probar dicha interfaz mediante la utilización de cualquier navegador internet es: `http://gplsi.dlsi.ua.es/wsd`.

Contexto: *plane, air*
Palabras no desambiguadas: *plane*
Sentidos finales: *plane#1, plane#2, plane#3, plane#4, plane#5.*

Para Plane#1: → Peso = **1**

airplane, aeroplane, plane -- (an aircraft that has fixed a wing and is powered by propellers or jets; "the flight was delayed due to trouble with the airplane")
=> aircraft -- (a vehicle that can fly)
=> craft -- (a vehicle designed for navigation in or on water or **air** or through outer space)
=> vehicle -- (a conveyance that transports people or objects)
=> conveyance, transport -- (something that serves as a means of transportation)
=> instrumentality, instrumentation -- (an artifact (or system of artifacts) that is instrumental in accomplishing some end)
=> artifact, artefact -- (a man-made object)
=> object, physical object -- (a physical (tangible and visible) entity; "it was full of rackets, balls and other objects")
=> entity, something -- (anything having existence (living or nonliving))

Para Plane#2: → Peso = **0**

plane, sheet -- ((mathematics) an unbounded two-dimensional shape; "we will refer to the plane of the graph as the X-Y plane"; "any line joining two points on a plane lies wholly on that plane")
=> shape, form -- (the spatial arrangement of something as distinct from its substance; "geometry is the mathematical science of shape")
=> attribute -- (an abstraction belonging to or characteristic of an entity)
=> abstraction -- (a general concept formed by extracting common features from specific examples)

Para Plane#3: → Peso = **0**

plane -- (a level of existence or development; "he lived on a worldly plane")
=> degree, level, stage, point -- (a specific identifiable position in a continuum or series or especially in a process; "a remarkable degree of frankness"; "at what stage are the social sciences?")
=> state -- (the way something is with respect to its main attributes; "the current state of knowledge"; "his state of health"; "in a weak financial state")

Figura 3.17. Aplicación de la Heurística de la Glosa del Hiperónimo

Contexto: *action, court, ward, cost, servant, criticism, jury*
Palabras no desambiguadas: *cost*
Sentidos finales: *cost#2, cost#3*

Para Cost#1 → Peso = (1/5)+(1/6)+(1/6)+(1/3)+(1/3) = **1.2**

=> expense, disbursal, disbursement -- (amounts paid for goods and services that may be currently tax deductible (as opposed to capital expenditures))
=> business expense, trade expense -- (ordinary and necessary expenses incurred in a taxpayer's business or trade)
=> lobbying expense -- (expenses incurred in promoting or evaluating legislation; "many lobbying expenses are deductible by a taxpayer")
.....
=> relief -- ((law) redress awarded by a **court**; "was the relief supposed to be protection from future harm or compensation for past injury?")
=> actual damages, compensatory damages, general damages -- ((law) compensation for losses that can readily be proven to have occurred and for which the injured party has the right to be compensated)
=> nominal damages -- ((law) a trivial sum (usually $1.00) awarded as recognition that a legal injury was sustained (as for technical violations of a contract))
=> punitive damages, exemplary damages, smart money -- ((law) compensation in excess of actual damages (a form of punishment awarded in cases of malicious or willful misconduct))
=> double damages -- (twice the amount that a **court** would normally find the injured party entitled to)
=> treble damages -- (three times the amount that a **court** would normally find the injured party entitled to)
=> atonement, expiation, satisfaction -- (compensation for a wrong; "we were unable to get satisfaction from the local store")
=> counterbalance, offset -- (a compensating equivalent)
=> reparation -- (compensation exacted from a defeated nation by the victors)
=> refund -- (money returned to a payer)
=> rebate, discount -- (a refund of some fraction of the amount paid)
=> rent-rebate -- ((British) a rebate on rent given by a local government authority)
=> conscience money -- (payment made voluntarily to reduce guilt over dishonest dealings)
=> support payment -- (a payment made by one person for the support of another)
=> palimony -- (support paid by one half of an unmarried partnership after the relationship ends)
=> alimony, maintenance -- (**court**-ordered support paid by one spouse to another after they are separated)
=> child support -- (**court**-ordered support paid by one spouse to the other who has custody of the children after the parents are separated)
=> reward -- (payment made in return for a service rendered)
.....

Para Cost#2→ Peso=**0**

=> average cost -- (total cost for all units bought (or produced) divided by the number of units)
=> marginal cost, incremental cost, differential cost -- (the increase or decrease in costs as a result of one more or one less unit of output)
=> expensiveness -- (the quality of being high-priced)
=> costliness, dearness -- (the quality possessed by something with a great price or value)
=> lavishness, luxury, sumptuosity, sumptuousness -- (the quality possessed by something that is excessively expensive)
=> assessment -- (the market value set on assets)
=> tax assessment -- (the value set on taxable property)
=> inexpensiveness -- (the quality of being affordable)
=> reasonableness, moderateness, modestness -- (the property of being moderate in price; "the store is famous for the reasonableness of its prices")
=> bargain rate, cheapness, cut rate, cut price -- (a price below the standard price)

Figura 3.18. Aplicación de la Heurística de la Glosa del Hipónimo

Heurística del Hiperónimo

```
Peso = 0. // Cada sentido tiene inicialmente un peso 0.
Para todas las palabras no desambiguadas completamente
   Para todos los sentidos
      Obtener synsets hiperónimos.
   Fin para
      Para el resto de palabras del contexto
         Buscar en synsets de los hiperónimos.
         Si aparece entonces
            peso = peso + (N nivel / N niveles total)
         Fin si
      Fin para
   Coger el sentido con mayor peso como solución.
Fin para
```

Figura 3.19. Algoritmo heurística hiperónimo

En primer lugar se describirá la arquitectura utilizada para desarrollar esta interfaz, para posteriormente describir las operaciones que se pueden realizar en la misma.

Arquitectura de la interfaz. Inicialmente el usuario introduce un grupo de palabras a la interfaz web, que se envían al servidor web, para activar un proceso que comprueba y estructura adecuadamente la información introducida. Este proceso WSD realiza la desambiguación de los nombres introducidos mediante la aplicación del método de Marcas de Especificidad y el uso de la base de datos léxica WordNet. Cuando el módulo WSD termina su proceso, se activa otro proceso con el objetivo de estructurar la información desambiguada y enviarla a la interfaz web para que los usuarios puedan ver el resultado de la desambiguación.

Operaciones de la interfaz. La apariencia externa de la interfaz, que se ilustra en la Figura 3.25, consiste en dos ventanas de texto, que sirven para introducir los nombres y para mostrar los

Heurística de Definición

```
Peso = 0.  // Cada sentido tiene inicialmente un peso 0.
Para todas las palabras no desambiguadas completamente
   Para todos los sentidos
      Para el resto de palabras del contexto
         Buscar en la definición del sentido.
         Si aparece  entonces
            peso = peso + 1
         Fin si
      Fin para
   Fin para
   Coger el sentido con mayor peso como solución.
Fin Para
```

Figura 3.20. Algoritmo heurística definición

sentidos de esas palabras y en dos comandos de ejecución, que sirven para realizar las dos operaciones que se van a describir a continuación.

- *Proceso WSD.* Este comando permite a los usuarios ejecutar el algoritmo que resuelve la ambigüedad léxica propuesto en esta Tesis y explicado en las secciones anteriores. La entrada al algoritmo son los nombres que se introducen en la ventana de texto que aparece a la izquierda de la interfaz, y que se llama "Nombres a Desambiguar". El resultado de la desambiguación se muestra en la ventana de texto que aparece a la derecha de la interfaz, y que se llama "Sentidos de WordNet". La información que muestra esta ventana se divide en cuatro columnas:
  - La primera columna muestra el número de *synset* de WordNet correspondiente al sentido elegido.
  - La segunda columna muestra el nombre que se quiere desambiguar.

Heurística del Hipónimo

```
Peso = 0. // Cada sentido tiene inicialmente un peso 0.
  Para todas las palabras no desambiguadas completamente
    Para todos los sentidos
      Coger synsets hipónimos.
    Fin para
      Para el resto de palabras del contexto
        Buscar en synset de los hipónimos.
        Si aparece entonces
          peso = peso + 1 / nivel de profundidad
        Fin si
      Fin para
    Coger el sentido con mayor peso como solución.
  Fin para
```

Figura 3.21. Algoritmo heurística hipónimo

- La tercera muestra el número de sentido proporcionado por WordNet correspondiente al sentido elegido entre todos los posibles.
- Y la última columna muestra la glosa proporcionada por WordNet correspondiente al sentido elegido.

- *Limpiar*. Este comando permite borrar la información que aparece en ambas ventanas de texto para una desambiguación realizada y actualiza el algoritmo para realizar otro ejemplo nuevo.

Cuando el algoritmo no es capaz de seleccionar un único sentido para una palabra, es decir devuelve varios sentidos para dicha palabra, la interfaz muestra un asterisco (*) para cada uno de los sentidos elegidos.

En resumen, en este capítulo se ha presentado el método de Marcas de Especificidad propuesto en esta Tesis para la desambiguación léxica de los nombres en cualquier lengua que tenga

Heurística de la Glosa del Hiperónimo

```
Peso = 0. // Cada sentido tiene inicialmente un peso 0.
  Para todas las palabras no desambiguadas completamente
    Para todos sus sentidos
      Coger hiperónimos.
    Fin para
      Para el resto de palabras del contexto
        Buscar en la definición de los hiperónimos.
        Si aparece entonces
          peso = peso + 1
        Fin si
      Fin para
    Coger el sentido con mayor peso como solución.
  Fin para
```

Figura 3.22. Algoritmo heurística glosa hiperónimo

un WordNet específico. También se ha presentado la descripción formal del algoritmo que realiza la desambiguación y finalmente se ha diseñado e implementado dicho algoritmo mediante una interfaz web con el objetivo de comprobar el funcionamiento del método de Marcas de Especificidad propuesto en esta Tesis.

Heurística de la Glosa del Hipónimo

```
Peso = 0. // Cada sentido tiene inicialmente un peso 0.
  Para todas las palabras no desambiguadas completamente
    Para todos sus sentidos
      Coger hipónimos.
    Fin para
      Para el resto de palabras del contexto
        Buscar en la definición del hipónimo.
        Si aparece entonces
          peso = peso + 1 / nivel de profundidad
        Fin si
      Fin para
    Coger el sentido con mayor peso como solución.
  Fin para
```

Figura 3.23. Algoritmo heurística glosa hipónimo

Heurística de Marca de Especificidad Común

```
Para todas las palabras no desambiguadas completamente
  Para todos los sentidos sin desambiguar
    Obtener hiperónimos.
    Buscar synset común a todos
      Si aparece entonces
        Elegir los sentidos
      Fin si
  Fin para
Fin para
```

Figura 3.24. Algoritmo heurística marca de especificidad común

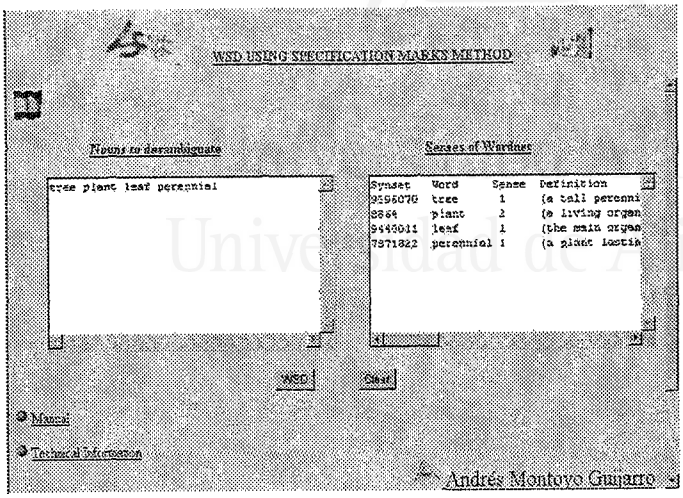

Figura 3.25. Interfaz web

Volver al índice/Tornar a l'índex

4. Experimentación

En este capítulo se presenta la evaluación realizada con el método de Marcas de Especificidad presentado en el capítulo anterior. Para llevar a cabo este trabajo se ha realizado inicialmente una fase de adaptación o validación, con el objetivo de mejorar el método hasta conseguir una versión final, para posteriormente realizar la fase de evaluación, la cual consiste en presentar resultados de funcionamiento del método de Marcas de Especificidad. Para la realización de ambas fases, en primer lugar, se presenta una introducción al proceso de evaluación de los sistemas de WSD. En segundo lugar, se realizará una descripción detallada del entorno experimental, presentando los recursos empleados y un análisis de los distintos parámetros empleados en los experimentos. Y finalmente, se muestran los experimentos realizados con el objetivo de mejorar el método hasta conseguir una versión final.

4.1 Introducción a la evaluación de WSD

La evaluación de los sistemas de WSD ha sido muy heterogénea debido a los tipos de corpus, conjunto de palabras, tamaños de los corpus, colecciones de sentidos, métodos de evaluación (artificiales, etc) utilizados para ello. Por lo tanto, la comparación entre ellos siempre ha sido difícil hasta que aparecieron los corpus anotados semánticamente como SemCor o DSO. Así, los sistemas WSD comparan sus resultados con las etiquetadas proporcionadas por estos corpus. Además, hay una competición científica sobre WSD llamada SENSEVAL, cuyo propósito es evaluar la potencia y la debilidad de los sistemas WSD con respecto a diferentes palabras, diferentes variedades de lengua y diferentes lenguas. Para

más información sobre SENSEVAL ir al apartado 2.5., y más concretamente al 2.5.1.

Algunos autores como Gale *et al.* (1992a) presentan una extensa discusión del problema de la evaluación de los métodos de WSD y destacan que algunas palabras son difíciles de desambiguar para algunos métodos de WSD, pero que hay otras que resultan fáciles. Otros autores como Resnik y Yarowsky (1997), disertan sobre la evaluación con la afirmación de que está lejos de ser estandarizada, ya que depende, tanto de los recursos lingüísticos que se utilicen como de la granularidad y el número de significados establecidos. Concluyen, con la propuesta de un conjunto de datos de evaluación o corpus estándar (*Gold Standard Datasets* (Kilgarriff, 1998)). Otros autores como Escudero *et al.* (Escudero et al., 2000a) presentan un trabajo sobre la evaluación cuando se cambia de dominio o de genero literario.

Salton (1989) expone que en la evaluación de un método se pueden diferenciar aspectos relacionados con la efectividad y aspectos relacionados con la eficiencia. La efectividad de un método se centra en la precisión con que éste resuelve la ambigüedad léxica. Los aspectos relacionados con la eficiencia se centran en cuestiones referentes al tiempo consumido por el sistema en el proceso de entrenamiento, al tiempo de ejecución, al tiempo consumido en el análisis de los documentos a desambiguar, etc. En el presente capítulo, nos centramos exclusivamente en el estudio de aspectos relacionados con la efectividad.

La efectividad de un método se puede medir mediante dos tipos de evaluación: *Evaluación directa* e *indirecta* (Ureña, 1998).

4.1.1 Evaluación directa

La evaluación directa mide la efectividad en la asignación de los significados correctos a las palabras a desambiguar. Esta evaluación es fundamental para cuantificar la calidad de los métodos de desambiguación. Cuando se aplica este tipo de evaluación surgen distintos problemas como:

- La escasez de colecciones de evaluación, debido a la dificultad y al coste que conlleva etiquetar semánticamente un corpus. Aunque últimamente están apareciendo cada vez más.
- La inconsistencia en el etiquetado de las colecciones de evaluación, ya que distintas personas pueden asignar diferentes significados a la misma palabra en el mismo contexto.
- La falta de acuerdo en la elección de definiciones de las palabras, debido a que diferentes diccionarios suelen proporcionar distintos conjuntos de sentidos para la misma palabra.
- La falta de criterios o acuerdos en las métricas utilizadas, debido a que los autores presentan diferentes formas de medir la efectividad de la desambiguación.

SENSEVAL[1] intenta resolver los problemas comentados anteriormente. En primer lugar, se define una colección de evaluación única para todos los participantes en la competición, y paralelamente un conjunto de significados adaptados a la colección. En segundo lugar, todos los sistemas etiquetarán los sentidos de las palabras a partir de los sentidos suministrados por HECTOR para SENSEVAL-1 y WordNet para SENSEVAL-2. Por último, las métricas utilizadas son "cobertura" (*recall*) y "precisión" (*precision*).

A continuación se describirán detalladamente estas dos métricas. Tanto "cobertura" como "precisión" son dos índices relacionados con la efectividad de los sistemas de desambiguación. "Precisión" se describe formalmente mediante la fórmula 4.1, definiéndose como el ratio entre los sentidos desambiguados correctamente y el número total de sentidos contestados.

$$Precision = \frac{sentidos\ desambiguados\ correctamente}{N^{o}\ sentidos\ contestados} \quad (4.1)$$

"Cobertura" se define como el ratio entre los sentidos desambiguados correctamente y el número total de sentidos, y se describe formalmente mediante la fórmula 4.2.

[1] SENSEVAL es una competición con el objetivo de evaluar sus sistemas de desambiguación (Kilgarriff, 1998). Esta competición propone una colección de textos de evaluación, un conjunto de definiciones y una serie de métricas concretas, y logra una alta consistencia en el etiquetado de la colección.

$$Cobertura = \frac{sentidos\ desambiguados\ correctamente}{N^o\ total\ sentidos} \quad (4.2)$$

También, se puede presentar un tercer índice de menor importancia denominado "cobertura absoluta" (*Coverage*), el cual se define como el ratio entre el número total de sentidos contestados y el número total de sentidos. Su descripción formal corresponde a la fórmula 4.3.

$$Cobertura\ absoluta = \frac{N^o\ total\ sentidos\ contestados}{N^o\ total\ sentidos} \quad (4.3)$$

4.1.2 Evaluación indirecta

La desambiguación de sentidos puede verse como una tarea intermedia. Por ello la evaluación indirecta mide la efectividad de la desambiguación respecto a otra tarea final (recuperación de información, traducción automática, categorización de documentos, etc) a la que se aplica, en función del método de desambiguación empleado. Cada tarea se evalúa de una manera distinta, con sus propias métricas y colecciones. Esta evaluación es fundamental para cuantificar la calidad de los distintos sistemas de desambiguación sobre la tarea a la que se aplica. En el capítulo 5, se describirá el procedimiento de aplicación de la desambiguación a la tarea concreta del enriquecimiento de WordNet mediante un sistema de clasificación denominado IPTC.

4.2 Descripción del entorno experimental

A continuación se presenta el entorno experimental utilizado con el objetivo de estudiar la efectividad de nuestro método. En primer lugar, se describirán los recursos empleados en los distintos experimentos. Finalmente se analizarán las distintas ventanas contextuales empleadas en los experimentos para la desambiguación de los distintos nombres que aparecen en ella.

Hay que aclarar que todos los experimentos se basan en las medidas suministradas por los índices cobertura absoluta, cobertura y precisión.

4.2.1 Recursos empleados

Los experimentos se realizan a partir del uso de los siguientes recursos:

- **Corpus de evaluación.** Se han seleccionado dos corpus de entrada para medir la efectividad de nuestro método de Marcas de Especificidad: *SemCor* y *Microsoft Encarta 98 Encyclopedia Deluxe*. Aunque también se utilizaron los recursos y corpus propuestos por el comité de SENSEVAL-2.
  - *SemCor* (G. Miller y Bunker, 1993) es parte del *Brown Corpus*, etiquetado manualmente con los sentidos de las palabras definidas en WordNet. Este corpus etiquetado semánticamente está disponible gratuitamente en la Web[2] para la comunidad investigadora.
    SemCor se ha construido a partir de dos corpus de textos: por un lado, incluye 103 pasajes del corpus estándar *Presen-Day* editado por American English (el Brown Corpus) de 1.014.312 palabras y por otro, del texto completo de la novela de Stephen Crane *The Red Badge of Courage* con un total de 45.600 palabras. Consta de 500 pasajes de 2000 palabras cada uno, extraídos de ediciones de documentos contemporaneos. Fue diseñado como una colección de textos heterogéneos y equilibrada a través de diferentes estilos y géneros literarios, tratando temas políticos, científicos, literarios, deportivos, musicales, etc.
    En definitiva, SemCor consta de unas 250.000 palabras, donde todas ellas se etiquetaron manualmente con los sentidos de WordNet. Las estadísticas de Semcor se muestran en la Tabla 4.1.
    Las palabras de un documento tienen distintas frecuencias de aparición dentro de una colección. Así, si se realiza un estudio de esta cualidad en la colección SemCor, en relación con las palabras ambiguas, se obtienen los resultados que se muestran en la Tabla 4.2. En esta tabla se muestran los porcentajes de ocurrencias con que aparecen los distintos sentidos, con

[2] `http://www.cogsci.princeton.edu/~ wn/`

SemCor				
	Brown1	Brown2	Brownv	Total
Palabras etiquetadas	198.796	160.936	316.814	676.546
Palabras con punteros semánticos a WN	106.639	86.000	41.497	234.136
Palabras etiquetadas con sentidos múltiples	115	551	37	703
Etiquetas semánticas	106.725	86.414	41.525	234.664
Palabras no-etiquetadas	92.154	74.936	135.684	302.774
Punteros semánticos a nombres	48.835	39.477	0	88.312
Punteros semánticos a verbos	26.686	21.804	41.525	90.015
Punteros semánticos a adjetivos	9.886	7.539	0	17.425
Punteros semánticos a adverbios	11.347	9.245	0	20.592
Punteros a nombres propios	5.602	4.075	0	9.684
Sentidos apuntados por nombres	11.399	9.546	0	20.945
Sentidos apuntados por verbos	5.334	4.790	6.520	16.644
Sentidos apuntados por adjetivos	1.754	1.463	0	3.217
Sentidos apuntados por adverbios	1.455	1.377	0	2.832

Tabla 4.1. Estadísticas SemCor.

paréntesis se indica el sentido más frecuente con el que aparece en la colección y en la última columna (entre llaves) se muestra el porcentaje que resultaría si los sentidos apareciesen en igual cantidad.

- En primer lugar, para Microsoft Encarta 98 Encyclopedia Deluxe se seleccionó una pequeña parte de la misma, para posteriormente de forma manual etiquetar cada una de las palabras con los sentidos definidos en WordNet. Como se puede apreciar esta parte del corpus fue realizada de manera artesanal, ya que no se disponía de otro corpus etiquetado y era aconsejable en los experimentos probar más de un corpus.
- También, hay que destacar que se participó en la competición desambiguación denominada SENSEVAL-2 y se utilizaron los recursos y corpus propuestos por la misma.
- Por último, para realizar otros experimentos más específicos se seleccionaron un grupo determinado de palabras ambiguas

N° Sentidos	Frecuencia	Sentido más común (%)		
1	53442	100 %	(1)	{100}
2	28791	77 %	(1)	{50}
3	25134	69 %	(1)	{33}
4	17265	63 %	(1)	{25}
5	11393	57 %	(1)	{20}
6	9334	54 %	(1)	{17}
7	5943	52 %	(1)	{14}
8	5543	53 %	(1)	{13}
9	3521	53 %	(1)	{11}
10	11137	63 %	(1)	{10}
11	1412	50 %	(1)	{9}
12	1232	45 %	(1)	{8}
13	2053	29 %	(1)	{8}
14	794	30 %	(1)	{7}
15	506	37 %	(1)	{7}
16	601	45 %	(1)	{6}
17	555	54 %	(1)	{6}
18	131	26 %	(1)	{6}
19	922	44 %	(1)	{5}
20	1		(1)	{5}
21	1714	46 %	(1)	{5}
22	1		(1)	{5}
23	126	16 %	(2)	{4}
24	1		(1)	{4}
25	1		(1)	{4}
26	1		(1)	{4}
27	1	27 %	(1)	{4}
28	1		(1)	{4}
29	758	35 %	(1)	{3}
30	1		(1)	{3}
31	1		(1)	{3}
32	152	22 %	(3)	{3}
33	1		(1)	{3}
34	1		(1)	{3}
35	356	10 %	(3)	{3}

Tabla 4.2. Porcentaje de ocurrencias de SemCor.

con el objetivo de asignar el mejor sentido para cada una de ellas.

- **Base de datos léxica.** Se ha seleccionado para todos los experimentos la base de conocimiento WordNet, presentada en el capítulo tres.

4.2.2 Tamaño de la ventana contextual

Uno de los objetivos de los presentes experimentos fue decidir entre diferentes unidades contextuales: frase, párrafo o bien otros tamaños de ventanas contextuales. Para ello, se seleccionaron aleatoriamente textos de los siguientes documentos de *SemCor* : *br-r09, br-m01, br-k10, br-j22, br-e22, br-f44, br-j41 y br-n20.* Se realizaron diferentes experimentos con diferentes tamaños de ventanas contextuales (frase, 10 a 35 palabras, etc) a partir de los textos anteriormente comentados como palabras de entrada al método de Marcas de Especificidad. Finalmente y después de realizar diferentes experimentos se seleccionó aquella ventana contextual que optimiza nuestro método.

4.3 Trabajo experimental

En esta sección se presenta el trabajo experimental realizado con el objetivo de estudiar la efectividad de nuestro método de Marcas de Especificidad. En primer lugar, se describen los tres experimentos realizados con el objetivo de perfeccionar y ajustar el método de Marcas de Especificidad para obtener los mejores resultados de desambiguación. El experimento 1 consiste en comprobar el funcionamiento del método cuando se aplica solamente la noción de Marca de Especificidad. El experimento 2 consiste en comprobar que al complementar el método con el conjunto de heurísticas presentadas en el capítulo anterior, estas aportan mejores porcentajes de desambiguación y por lo tanto mejoran el método. Y el experimento 3 consiste en analizar y definir la ventana óptima de contexto para obtener la mejor desambiguación.

En segundo lugar se presentan los dos experimentos efectuados con el objetivo de realizar una comparación *directa* e *indirecta* del

método de Marcas de Especificidad con otros métodos de WSD. En el experimento 4 se realiza una comparación con métodos basados en el conocimiento, es decir métodos WSD que pertenecen a la misma clasificación que el propuesto en esta Tesis. En el experimento 5 se realiza una comparación entre el método propuesto y uno basado en corpus, concretamente en un modelo probabilístico que utiliza el principio de Máxima Entropía.

Finalmente, se describe la evaluación final del método de Marcas de Especificidad con dos experimentos. El experimento 6 describe la participación en la competición de sistemas de WSD SENSEVAL-2 para los nombres seleccionados por el comité en la tarea de "lexical sample" tanto para inglés como para español. Y el experimento 7 y último consiste en evaluar al método aplicando las heurísticas en cascada (secuencialmente) y aplicándolas independientemente unas de otras, y en ambos casos sobre todos los documentos del corpus *SemCor*.

4.3.1 Trabajo experimental para el ajuste del método

Experimento 1: Método de Marcas de Especificidad.

Objetivo. El objetivo de este experimento es comprobar que el método de Marcas de Especificidad obtiene resultados satisfactorios de desambiguación cuando se aplica sin heurísticas, es decir se aplica solamente la noción de Marcas de Especificidad.

Descripción del recurso. La evaluación del método Marcas de Especificidad se ha realizado sobre textos del corpus *Semantic Concordance (Semcor)* y de la enciclopedia electrónica *Microsoft Encarta 98 Encyclopedia Deluxe*.

Los textos a desambiguar han sido escogidos al azar teniendo en cuenta que la ventana contextual es de una frase. El total de texto escogido es de 100 frases y 619 nombres para el *Semcor* y 100 frases y 697 nombres para la enciclopedia. Estos nombres elegidos contienen tanto palabras monosémicas como polisémicas. Se hace así porque, al tratarse de un modelo que se basa en el conocimiento suministrado por la base de conocimiento WordNet, las palabras monosémicas ayudan a desambiguar a las polisémicas. Para *Semcor* se han obtenido 111 nombres monosémicos y para la

enciclopedia 119. Además, para *Semcor* se ha obtenido el sentido más frecuente de WordNet (es decir, sentido #1) en 411 veces y 446 para la enciclopedia. Por lo tanto y a partir de estos resultados el grado de acierto del sentido más frecuente sería de 66,39 % en el *Semcor* y del 63,9 % en la enciclopedia.

Resultados. Los resultados de desambiguación, tanto de las polisémicas como de las monosémicas, obtenidos para el corpus *SemCor* son de 325 palabras desambiguadas correctamente, es decir el sistema ha conseguido obtener el mejor sentido de cada una de las palabras en la frase. En cuanto a las palabras que no se han desambiguado correctamente corresponden 173. Por último, el sistema ha devuelto 121 palabras sin desambiguar, es decir no ha sido capaz de asignarle un sentido. Por lo tanto, los porcentajes de desambiguación obtenidos al aplicar el sistema propuesto han sido los siguientes; el 52,5 % de las palabras han sido desambiguadas correctamente, el 28 % incorrectamente y el 19,5 % no han podido desambiguarse, bien porque esas palabras no están en WordNet o porque el método es incapaz de obtener un sentido para ellas.

Los resultados de desambiguación obtenidos para el corpus *Microsoft Encarta 98*, tanto de las polisémicas como de las monosémicas, son de 385 palabras desambiguadas correctamente, 186 incorrectamente y 126 sin desambiguar. Por lo tanto, los porcentajes de desambiguación obtenidos al aplicar el sistema propuesto han sido los siguientes; el 55,2 % de las palabras han sido desambiguadas correctamente, el 26,7 % incorrectamente y el 18,1 % no han podido desambiguarse por los mismos motivos explicados anteriormente.

Un resumen de los resultados sin diferenciar monosémicas de polisémicas se muestra en la Tabla 4.3.

Sin embargo, si se tienen en cuenta solo los nombres polisémicos, los resultados de desambiguación obtenidos para el corpus *SemCor* son de 214 palabras desambiguadas correctamente, es decir el sistema ha conseguido obtener el mejor sentido de cada una de las palabras en la frase. En cuanto a las palabras que no se han desambiguado correctamente corresponden 173. Por último, el sistema ha devuelto 121 palabras sin desambiguar, es decir no

	SemCor			Encarta 98		
	Bien	*Mal*	*Sin Des.*	*Bien*	*Mal*	*Sin des.*
N° Palabras	325	173	121	385	186	126
%	52,5%	28%	19,5%	55,2%	26,7%	18,1%

Tabla 4.3. Resultados al aplicar el Sistema Marcas Especificidad sin Heurísticas (Monosémicas y Polisémicas).

ha sido capaz de asignarle un sentido. Por lo tanto, los porcentajes de desambiguación obtenidos al aplicar el sistema propuesto han sido los siguientes; el 42,13 % de las palabras han sido desambiguadas correctamente, el 34,1 % incorrectamente y el 23,8 % no han podido desambiguarse, bien porque esas palabras no están en WordNet o porque el método es incapaz de obtener un sentido para ellas.

En el caso de solo nombres polisémicos, los resultados de desambiguación obtenidos para el corpus *Microsoft Encarta 98* son de 266 palabras desambiguadas correctamente, 186 incorrectamente y 126 sin desambiguar. Por lo tanto, los porcentajes de desambiguación obtenidos al aplicar el sistema propuesto han sido los siguientes; el 46,02 % de las palabras han sido desambiguadas correctamente, el 32,18 % incorrectamente y el 21,8 % no han podido desambiguarse por los mismos motivos explicados anteriormente.

Un resumen de los resultados con solo nombres polisémicos se muestra en la Tabla 4.4.

	SemCor			Encarta 98		
	Bien	*Mal*	*Sin Des.*	*Bien*	*Mal*	*Sin des.*
N° Palabras	214	173	121	266	186	126
%	42,13%	34,1%	23,8%	46,02%	32,18%	21,8%

Tabla 4.4. Resultados al aplicar el Sistema Marcas Especificidad sin Heurísticas (Solo polisémicas).

Los resultados anteriores no están expresados en las medidas estándares ("precisión", "cobertura" y "cobertura absoluta") utilizadas para evaluar los sistemas WSD. Por tal motivo, la Tabla 4.5 muestra los resultados con dichos valores, sin diferenciar los

nombres polisémicos y los monosémicos, obtenidos al aplicar las fórmulas 4.1, 4.2 y 4.3, respectivamente.

%	Cobertura absoluta	Precisión	Cobertura
Método Marcas Especificidad	81,2 %	66,4 %	54 %

Tabla 4.5. Medidas del método de Marcas Especificidad sin Heurísticas (Monosémicas y Polisémicas).

Los resultados de las mismas medidas de 4.1, 4.2 y 4.3 pero solo con nombres polisémicos se muestra en la Tabla 4.6.

%	Cobertura absoluta	Precisión	Cobertura
Método Marcas Especificidad	77,3 %	57,21 %	44,2 %

Tabla 4.6. Medidas del método de Marcas Especificidad sin Heurísticas (solo polisémicas).

Discusión. Como puede observarse en las Tabla 4.3 y 4.4, los resultados obtenidos en la desambiguación de las palabras utilizando el sistema sin heurísticas no han sido del todo satisfactorios, ya que apenas sobrepasan el 50 % de aciertos.

El porcentaje obtenido de palabras sin desambiguar es bastante alto y en la mayoría de los casos, el resultado correcto está incluido entre esos sentidos sin desambiguar. Al observar los sentidos finales obtenidos, se llegó a la conclusión de que en muchos casos, el sistema no era capaz de dar un único sentido como resultado por no tener suficiente información para desambiguar la palabra. Por esta razón, en lugar de dar un sentido aleatorio o el más frecuente en WordNet, se decidió subir de nivel en la subjerarquía de hiperónimos hasta conseguir aquel synset común a esos sentidos, y así poder dar alguna información sobre la desambiguación de la palabra aunque esta fuese más general, pero no errónea.

La mayor parte de las relaciones semánticas representadas en WordNet son clase/subclase (hiperónimo/hipónimo), por ello, muchas de las relaciones semánticas que aparecen entre las palabras de los textos no pueden ser capturadas por este recurso. Así, se decidió añadir un peso a aquellos sentidos en los que en

su subjerarquía apareciese alguna palabra del resto de las palabras pertenecientes al conjunto de entrada. También ese peso se incrementó en aquellos casos que apareciese también alguna de las palabras del conjunto de entrada en las definiciones de los sentidos.

De estos estudios realizados con el objetivo de mejorar los resultados obtenidos por el algoritmo surgieron las heurísticas de Marcas de Especificidad Común, Hiperónimo, Hipónimo, Glosa Hiperónimo, Glosa Hipónimo y Definición que en el capítulo anterior se explicaron detalladamente.

Experimento 2: Marcas de Especificidad con heurísticas.

Objetivo. El objetivo de este experimento fue evaluar el conjunto de heurísticas que complementan y mejoran el método de Marcas de Especificidad. Las heurísticas en concreto se han denominado: de Hiperónimo, Definición, Hipónimo, Glosa de Hiperónimos, Glosa de Hipónimos y Marcas de Especificidad Común. Con este experimento se puede comprobar la mejora que aportan estas heurísticas aplicadas en cascada (una detrás de la otra) al proceso de desambiguación.

Descripción del recurso. La evaluación del método Marcas de Especificidad se ha realizado sobre textos del corpus *Semantic Concordance (Semcor)* y de la enciclopedia electrónica *Microsoft Encarta 98 Encyclopedia Deluxe.*

Los textos a desambiguar han sido escogidos al azar teniendo en cuenta que la ventana contextual es de una frase. El total de texto escogido es de 619 nombres para el *Semcor* y 697 para la enciclopedia. Es decir, se ha escogido el mismo que en el experimento 1.

Resultados. A continuación, se describirán los resultados al aplicar cada una de las seis heurísticas al conjunto de textos seleccionados. En primer lugar se describirá el rendimiento obtenido al ir aplicando consecutivamente cada una de las heurísticas y en segundo lugar el obtenido en general por todas ellas.

Corpus SemCor. Si se observa la Tabla 4.3, y en particular los datos para el corpus *SemCor*, se obtenían 121 palabras sin desambiguar. Al aplicar el método con las seis heurísticas se han

conseguido desambiguar correcta o incorrectamente 112 palabras, por lo que se alcanza una "cobertura absoluta" del 92,56 % en el proceso de desambiguación. Y solamente 9 palabras no se pueden desambiguar debido a que no se encuentra ningún *synset* común. A continuación se presentan los resultados parciales para cada una de las heurísticas (H):

- **H1:** *Heurística de Hiperónimo.* Al aplicar la heurística del hiperónimo a esas 121 palabras polisémicas sin desambiguar, el 11,57 % se desambiguan (correcta e incorrectamente), obteniéndose un porcentaje de desambiguación correcta del 21,43 % sobre ese 11,57 %. La aplicación de esta heurística es muy prometedora, ya que con su aplicación se desambiguan muchos casos no contemplados por las relaciones de hiperónimia, hiponimia y sinonimia. La Tabla 4.7 muestra los resultados obtenidos cuando se aplica esta heurística.

	ME			ME + H1		
	Bien	*Mal*	*Sin Des.*	*Bien*	*Mal*	*Sin des.*
N° Palabras	325	173	121	328	184	107
%	52,5%	28%	19,5%	53%	29,7%	17,3%

Tabla 4.7. Heurística Hiperónimo

Los resultados de las medidas de precisión, cobertura y cobertura absoluta aplicando al método de Marcas de Especificidad con la heurística hiperónimo para nombres polisémicos, se muestra en la Tabla 4.8.

%	Cobertura absoluta	Precisión	Cobertura
Método Marcas Especificidad	82,7 %	64,06 %	53 %

Tabla 4.8. Medidas Marcas Especificidad con H1

- **H2:** *Heurística de Definición.* Al aplicar la heurística de la definición a las 107 palabras sin desambiguar, el 5,60 % se desambiguan (correcta e incorrectamente) obteniéndose un porcentaje de desambiguación correcta del 50 % sobre ese 5,60 %. La Ta-

bla 4.9 muestra los resultados obtenidos cuando se aplica esta heurística.

	ME + H1			ME + H1 + H2		
	Bien	*Mal*	*Sin Des.*	*Bien*	*Mal*	*Sin des.*
N° Palabras	328	184	107	331	187	101
%	53%	29,7%	17,3%	53,5%	30,2%	16,3%

Tabla 4.9. Heurística Definición

Los resultados de las medidas de precisión, cobertura y cobertura absoluta aplicando además de las otras heurísticas anteriores la heurística de definición para nombres polisémicos, se muestra en la Tabla 4.10.

%	Cobertura absoluta	Precisión	Cobertura
Método Marcas Especificidad	83,7 %	63,9 %	53,5 %

Tabla 4.10. Medidas Marcas Especificidad con H1 y H2

- **H3:** *Heurística de Hipónimo.* Al aplicar la heurística del hipónimo a las 101 palabras sin desambiguar, el 4,95 % se desambiguan (correcta e incorrectamente) obteniéndose un porcentaje de desambiguación correcta del 20 % sobre ese 4,95 %. La Tabla 4.11 muestra los resultados obtenidos cuando se aplica esta heurística.

	ME + H1 + H2			ME + H1 + H2 + H3		
	Bien	*Mal*	*Sin Des.*	*Bien*	*Mal*	*Sin des.*
N° Palabras	331	187	101	332	191	96
%	53,5%	30,2%	16,3%	53,6%	30,8%	15,5%

Tabla 4.11. Heurística Hipónimo

Los resultados de las medidas de precisión, cobertura y cobertura absoluta aplicando además de las otras heurísticas anteriores la heurística de hipónimo para nombres polisémicos, se muestra en la Tabla 4.12.

%	Cobertura absoluta	Precisión	Cobertura
Método Marcas Especificidad	84,5 %	63,5 %	53,6 %

Tabla 4.12. Medidas Marcas Especificidad con H1, H2 y H3

- **H4:** *Heurística Glosa Hiperónimo.* Al aplicar la heurística de glosa hiperónimo a las 96 palabras sin desambiguar, el 15,62 % se desambiguan (correcta e incorrectamente) obteniéndose un porcentaje de desambiguación correcta del 53,33 % sobre ese 15,62 %. La Tabla 4.13 muestra los resultados obtenidos cuando se aplica esta heurística.

	ME + H1 + H2 + H3			**ME + H1 + H2 + H3 + H4**		
	Bien	*Mal*	*Sin Des.*	*Bien*	*Mal*	*Sin des.*
N° Palabras	332	191	96	340	198	81
%	53,6%	30,8%	15,5%	54,9%	31,8%	13%

Tabla 4.13. Heurística Glosa Hiperónimo

Los resultados de las medidas de precisión, cobertura y cobertura absoluta aplicando además de las otras heurísticas anteriores la heurística de glosa hiperónimo para nombres polisémicos, se muestra en la Tabla 4.14.

%	Cobertura absoluta	Precisión	Cobertura
Método Marcas Especificidad	86,9 %	63,2 %	54,9 %

Tabla 4.14. Medidas Marcas Especificidad con H1, H2, H3 y H4

- **H5:** *Heurística Glosa Hipónimo.* Al aplicar la heurística de glosa hipónimo a las 81 palabras sin desambiguar, el 16,05 % se desambiguan (correcta e incorrectamente) obteniéndose un porcentaje de desambiguación correcta del 92,30 % sobre ese 16,05 %. La Tabla 4.15 muestra los resultados obtenidos cuando se aplica esta heurística.
  Los resultados de las medidas de precisión, cobertura y cobertura absoluta aplicando además de las otras heurísticas anteriores

	ME+H1+H2+H3+H4			ME+H1+H2+H3+H4+H5		
	Bien	*Mal*	*Sin Des.*	*Bien*	*Mal*	*Sin des.*
N° Palabras	340	198	81	352	199	68
%	54,9%	31,8%	13%	56,8%	32,1%	10,9%

Tabla 4.15. Heurística Glosa Hipónimo

la heurística de glosa hipónimo para nombres polisémicos, se muestra en la Tabla 4.16.

%	Cobertura absoluta	Precisión	Cobertura
Método Marcas Especificidad	89 %	63,9 %	56,9 %

Tabla 4.16. Medidas Marcas Especificidad con H1, H2, H3, H4 y H5

- **H6:** *Marca Especificidad Común.* Al aplicar la heurística de la marca de especificidad común de esas 68 palabras, el 86,76 % se desambiguan (correcta e incorrectamente) obteniéndose un porcentaje de desambiguación correcta del 98,31 % sobre ese 86,76 %. La Tabla 4.17 muestra los resultados obtenidos cuando se aplica esta heurística

	ME+H1+H2+H3+H4+H5			ME+H1+...+H6		
	Bien	*Mal*	*Sin Des.*	*Bien*	*Mal*	*Sin des.*
N° Palabras	352	199	68	410	200	9
%	56,8%	32,1%	10,9%	66,2%	32,3%	1,5%

Tabla 4.17. Heurística Marcas Especificidad Común

Los resultados de las medidas de precisión, cobertura y cobertura absoluta aplicando además de las otras heurísticas anteriores la heurística de Marca Especificidad Común para nombres polisémicos, se muestra en la Tabla 4.18.

%	Cobertura absoluta	Precisión	Cobertura
Método Marcas Especificidad	98,5 %	67,2 %	66,2 %

Tabla 4.18. Medidas Marcas Especificidad con H1, H2, H3, H4, H5 y H6

Corpus Encarta 98. Si se observa la Tabla 4.3, y en particular los datos para el corpus Encarta 98, quedaban 126 palabras polisémicas sin desambiguar. Al aplicar el método con las seis heurísticas se han conseguido desambiguar correcta o incorrectamente 118 palabras, por lo que se alcanza una "cobertura" del 93,65 % de desambiguación. Y solamente 8 palabras no se pueden desambiguar debido a que no se encuentra ningún *synset* común.

- **H1:** *Heurística Hiperónimo.* Al aplicar la heurística del hiperónimo de esas 126 palabras, el 9,5 % se desambiguan (correcta e incorrectamente) obteniéndose un porcentaje de desambiguación correcta del 66,67 % sobre ese 10,17 %. La Tabla 4.19 muestra los resultados obtenidos cuando se aplica esta heurística.

	ME			ME + H1		
	Bien	*Mal*	*Sin Des.*	*Bien*	*Mal*	*Sin des.*
N° Palabras	385	186	126	393	190	114
%	55,2%	26,7%	18,1%	56,4%	27,2%	16,4%

Tabla 4.19. Heurística Hiperónimo

Los resultados de las medidas de precisión, cobertura y cobertura absoluta aplicando al método de Marcas de Especificidad con la heurística hiperónimo para nombres polisémicos, se muestra en la Tabla 4.20.

%	Cobertura absoluta	Precisión	Cobertura
Método Marcas Especificidad	83,6 %	67,4 %	56,4 %

Tabla 4.20. Medidas Marcas Especificidad con H1

- **H2:** *Heurística Definición.* Al aplicar la heurística de la definición de esas 114 palabras, el 13,16 % se desambiguan (correcta e incorrectamente) obteniéndose un porcentaje de desambiguación correcta del 86,67 % sobre ese 13,16 %. La Tabla 4.21 muestra los resultados obtenidos cuando se aplica esta heurística.
  Los resultados de las medidas de precisión, cobertura y cobertura absoluta aplicando además de las otras heurísticas anteriores

	ME + H1			ME + H1 + H2		
	Bien	*Mal*	*Sin Des.*	*Bien*	*Mal*	*Sin des.*
N° Palabras	393	190	114	406	192	99
%	56,4%	27,2%	16,4%	58,3%	27,5%	14,2%

Tabla 4.21. Heurística Definición

la heurística de definición para nombres polisémicos, se muestra en la Tabla 4.22.

%	Cobertura absoluta	Precisión	Cobertura
Método Marcas Especificidad	85,8 %	67,9 %	58,25 %

Tabla 4.22. Medidas Marcas Especificidad con H1 y H2

- **H3:** *Heurística Hipónimo.* Al aplicar la heurística del hipónimo a las 99 palabras sin desambiguar, el 15,15 % se desambiguan (correcta e incorrectamente) obteniéndose un porcentaje de desambiguación correcta del 66,67 % sobre ese 15,15 %. La Tabla 4.23 muestra los resultados obtenidos cuando se aplica esta heurística.

	ME + H1 + H2			ME + H1 + H2 + H3		
	Bien	*Mal*	*Sin Des.*	*Bien*	*Mal*	*Sin des.*
N° Palabras	406	192	99	416	197	84
%	58,3%	27,5%	14,2%	59,7%	28,2%	12%

Tabla 4.23. Heurística Hipónimo

Los resultados de las medidas de precisión, cobertura y cobertura absoluta aplicando además de las otras heurísticas anteriores la heurística de hipónimo para nombres polisémicos, se muestra en la Tabla 4.24.

%	Cobertura absoluta	Precisión	Cobertura
Método Marcas Especificidad	87,9 %	67,9 %	59,7 %

Tabla 4.24. Medidas Marcas Especificidad con H1, H2 y H3

- **H4:** *Heurística Glosa Hiperónimo.* Al aplicar la heurística de glosa hiperónimo a las 84 palabras sin desambiguar, el 13,10 % se desambiguan (correcta e incorrectamente) obteniéndose un porcentaje de desambiguación correcta del 54,54 % sobre ese 13,10 % . La Tabla 4.25 muestra los resultados obtenidos cuando se aplica esta heurística.

	ME + H1 + H2 + H3			**ME + H1 + H2 + H3 + H4**		
	Bien	*Mal*	*Sin Des.*	*Bien*	*Mal*	*Sin des.*
N° Palabras	416	197	84	422	202	73
%	59,7%	28,2%	12%	60,5%	28,9%	10,4%

Tabla 4.25. Heurística Glosa Hiperónimo

Los resultados de las medidas de precisión, cobertura y cobertura absoluta aplicando además de las otras heurísticas anteriores la heurística de glosa hiperónimo para nombres polisémicos, se muestra en la Tabla 4.26.

%	Cobertura absoluta	Precisión	Cobertura
Método Marcas Especificidad	89,5 %	67,6 %	60,5 %

Tabla 4.26. Medidas Marcas Especificidad con H1, H2, H3 y H4

- **H5:** *Heurística Glosa Hipónimo.* Al aplicar la Heurística de Glosa Hipónimo a las 73 palabras sin desambiguar, el 31,50 % se desambiguan (correcta e incorrectamente) obteniéndose un porcentaje de desambiguación correcta del 56,52 % sobre ese 31,50 %. La Tabla 4.27 muestra los resultados obtenidos cuando se aplica esta heurística.

	ME+H1+H2+H3+H4			**ME+H1+H2+H3+H4+H5**		
	Bien	*Mal*	*Sin Des.*	*Bien*	*Mal*	*Sin des.*
N° Palabras	422	202	73	435	212	50
%	60,5%	28,9%	10,4%	62,4%	30,4%	7,1%

Tabla 4.27. Heurística Glosa Hipónimo

Los resultados de las medidas de precisión, cobertura y cobertura absoluta aplicando además de las otras heurísticas anteriores la heurística de glosa hipónimo para nombres polisémicos, se muestra en la Tabla 4.28.

%	Cobertura absoluta	Precisión	Cobertura
Método Marcas Especificidad	92,8 %	67,2 %	62,4 %

Tabla 4.28. Medidas Marcas Especificidad con H1, H2, H3, H4 y H5

- **H6:** *Heurística marca especificidad común.* Al aplicar la heurística de la marca de especificidad común de esas 50 palabras, el 84 % se desambiguan (correcta e incorrectamente) obteniéndose un porcentaje de desambiguación del 64,28 % sobre ese 84 %. La Tabla 4.29 muestra los resultados obtenidos cuando se aplica esta heurística

	ME+H1+H2+H3+H4+H5			**ME+H1+...+H6**		
	Bien	*Mal*	*Sin Des.*	*Bien*	*Mal*	*Sin des.*
N° Palabras	435	212	50	462	227	8
%	62,4%	30,4%	7,1%	66,3%	32,5%	1,1%

Tabla 4.29. Heurística de marcas especificidad común

Los resultados de las medidas de precisión, cobertura y cobertura absoluta aplicando además de las otras heurísticas anteriores la heurística de Marca Especificidad Común para nombres polisémicos, se muestra en la Tabla 4.30.

%	Cobertura absoluta	Precisión	Cobertura
Método Marcas Especificidad	98,8 %	67 %	66,3 %

Tabla 4.30. Medidas Marcas Especificidad con H1, H2, H3, H4, H5 y H6

Una vez expuestos los resultados parciales de la aplicación de cada una de las heurísticas propuestas, a continuación en la Tabla 4.31 se muestran los resultados totales del método de Marcas de Especificidad y las heurísticas.

Los resultados de desambiguación obtenidos para el corpus SemCor son de 410 palabras desambiguadas correctamente, es decir el método ha conseguido obtener el mejor sentido de cada una de las palabras en la frase. En cuanto a las palabras que no se han desambiguado correctamente corresponden 200. Por último, el sistema ha devuelto 9 palabras sin desambiguar, es decir no ha sido capaz de asignarle un sentido. Por lo tanto, los porcentajes de desambiguación obtenidos al aplicar el sistema propuesto han sido los siguientes; el 66,2 % de las palabras han sido desambiguadas correctamente, el 32,3 % incorrectamente y el 1,5 % no han podido desambiguarse, bien porque esas palabras no están en WordNet o porque el método es incapaz de obtener un único sentido para ellas.

Los resultados de desambiguación obtenidos para el corpus Microsoft Encarta 98 son de 462 palabras desambiguadas correctamente, 227 incorrectamente y 8 sin desambiguar. Por lo tanto, los porcentajes de desambiguación obtenidos al aplicar el sistema propuesto han sido los siguientes; el 66,3 % de las palabras han sido desambiguadas correctamente, el 32,5 % incorrectamente y el 1,1 % no han podido desambiguarse por los mismos motivos explicados anteriormente.

Un resumen de los resultados totales al aplicar conjuntamente el método de Marcas de Especificidad y las seis heurísticas se muestra en la Tabla 4.31 para SemCor.

	SEMCOR			
%	Bien	Mal	Sin Des.	Δ bien
ME Base	52,5%	28%	19,5%	-
ME Base+H1	53%	29,7%	17,3%	0,5
ME Base+H1+H2	53,5%	30,2%	16,3%	0,5
ME Base+H1+H2+H3	53,6%	30,8%	15,5%	0,1
ME Base+H1+H2+H3+H4	54,9%	31,8%	13%	1,3
ME Base+H1+H2+H3+H4+H5	56,8%	32,1%	10,9%	1,9
ME Base+H1+H2+H3+H4+H5+H6	66,2%	32,3%	1,5%	9,4

Tabla 4.31. Resultados del Sistema Marcas Especificidad con Heurísticas para SemCor

Y para Encarta se muestra en la Tabla 4.32 el resumen de resultados totales.

	ENCARTA			
%	Bien	Mal	Sin Des.	Δ bien
ME Base	55,2%	26,7%	18,1%	-
ME Base+H1	56,4%	27,2%	16,4%	1,2
ME Base+H1+H2	58,3%	27,5%	14,2%	3,9
ME Base+H1+H2+H3	59,7%	28,2%	12%	1,4
ME Base+H1+H2+H3+H4	60,5%	28,9%	10,4%	0,8
ME Base+H1+H2+H3+H4+H5	62,4%	30,4%	7,1%	1,9
ME Base+H1+H2+H3+H4+H5+H6	66,3%	32,5%	1,1%	3,9

Tabla 4.32. Resultados del Sistema Marcas Especificidad con Heurísticas para Encarta

Los resultados anteriores no están expresados en las medidas estándares ("precisión", "cobertura" y "cobertura absoluta") utilizadas para evaluar los sistemas WSD. Por tal motivo, la Tabla 4.33 muestra los resultados con dichos valores obtenidos al aplicar las fórmulas 4.1, 4.2 y 4.3, respectivamente.

%	Cobertura absoluta	Precisión	Cobertura
Método Marcas Especificidad	98,7 %	67,1 %	66,2 %

Tabla 4.33. Medidas del método de Marcas Especificidad con Heurísticas

Discusión. Como puede observarse en la Tabla 4.31, los resultados obtenidos en la desambiguación de las palabras al aplicar el método de Marcas de Especificidad junto con las seis heurísticas mejoran considerablemente los resultados previos mostrados en la Tabla 4.3, en donde solo se aplica el método sin heurísticas. Estos resultados podrían mejorar aún más si en vez de tener en cuenta las palabras de una frase solamente, se tuviesen en cuenta ventanas contextuales más grandes (por ejemplo de 10, 15, 30 palabras), ya que cuanto mayor sea el número de palabras de contexto mayor será el grado de conectividad esperado . La observación comentada en este párrafo se aclarará a continuación, ya que es el siguiente experimento a realizar.

Con este experimento se ha podido obtener la productividad de cada heurística cuando se aplican en cascada. Si se observan las tablas 4.31 y 4.32, hay una columna que indica el incremento de mejora que aporta cada heurística a partir de los nombres que no han podido desambiguarse por la heurística aplicada previamente.

Una conclusión obtenida de este experimento es que si las frases escogidas del SemCor tienen un contexto específico, es decir un contexto en el cual las palabras estén muy relacionadas semánticamente, el algoritmo funciona muy bien debido a que la distancia semántica de las palabras es pequeña. Sin embargo, si ocurre lo contrario, es decir si la distancia semántica de las palabras es grande cabe expresar que el algoritmo no funcione tan bien. Por este motivo, los resultados obtenidos han sido ligeramente mejores para el corpus Encarta 98, en donde las frases definen conceptos y por lo tanto la semejanza entre las palabras es mayor.

Experimento 3: Evaluación de la ventana de contexto.

Objetivo. El objetivo de este experimento es determinar la influencia del tamaño del contexto en el método de Marcas de Especificidad.

Descripción del recurso. La evaluación del método de Marcas de Especificidad conjuntamente con las seis heurísticas se ha realizado sobre textos del corpus Semantic Concordance (Semcor).

Tamaño del recurso. Los textos a desambiguar se seleccionaron aleatoriamente de los siguientes documentos del SemCor : *br-r09, br-m01, br-k10, br-j22, br-e22, br-f44, br-j41 y br-n20.* Para los ficheros anteriores se trataron 449, 402, 500, 484, 557, 550, 598 y 414 nombres respectivamente, siendo un total de 3954 nombres. Se tomaron diferentes tamaños de ventanas contextuales, desde 10 hasta 35 nombres a partir de los párrafos de los textos elegidos. El proceso seguido por este experimento fue el siguiente: para cada uno de los textos seleccionados, y para cada palabra a desambiguar se aplica el método para cada tamaño de la ventana. Por ejemplo, si el tamaño es de 10 palabras se eligen las 10 palabras contiguas y se desambiguan al mismo tiempo esas 10 palabras. Si no hay suficientes palabras para completar el contexto de la ventana deseada, entonces sobre ese contexto de palabras no se

realiza la desambiguación, así siempre se fuerza a que la ventana de contexto tenga la cantidad de palabras deseada.

Resultados. Los resultados de "precisión" (coincide con la "cobertura") , tanto de nombres monosémicos como polisémicos, obtenidos al aplicar el método sobre los textos seleccionados y para cada una de las ventanas contextuales indicadas, se muestran en la Tabla 4.34.

	Br-r09	*Br-m01*	*Br-k10*	*Br-j22*	*Br-e22*	*Br-f44*	*Br-j41*	*Br-n20*
10	50	60	70	20	20	20	30	90
15	53,3	53,3	73,3	33,3	53,3	33,3	46,6	86,6
20	60	45	65	20	45	25	45	60
25	64	48	68	20	56	16	60	56
30	60	40	56,7	23,3	50	30	43,3	53,3
35	54,3	40	45,7	25,7	57,1	31,4	45,7	54,2

Tabla 4.34. Resultados para las ventanas contextuales entre 10 y 35

La figura 4.1 muestra una gráfica indicando la *precision* obtenida de cada una de las ventanas contextuales sobre los textos elegidos del *SemCor*.

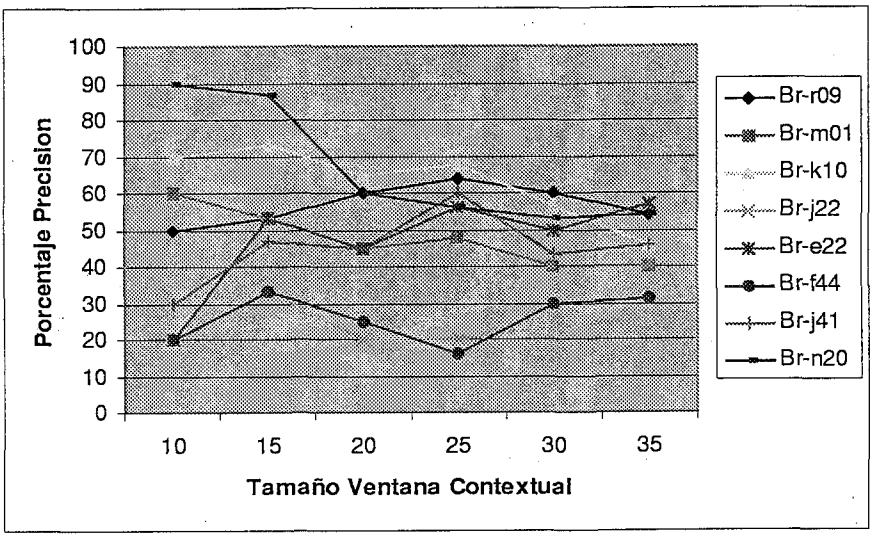

Figura 4.1. Gráfica de las ventanas contextuales entre 10 y 35

Una vez obtenidos estos resultados y con el objetivo de determinar la ventana óptima contextual, se decidió hacer la media aritmética ponderada respecto al número de nombres de cada fichero, de los valores obtenidos para cada uno de los textos sobre una misma ventana contextual. Los resultados se muestran en la Tabla 4.35, indicando que una ventana contextual con un intervalo aproximado a 15 palabras es el óptimo para desambiguar el sentido de las palabras mediante el método de Marcas de Especificidad. La figura 4.2 ilustra que ventana contextual es la óptima para el experimento realizado.

Ventana	10	15	20	25	30	35
%	45	54,13	45,63	48,5	44,58	44,26

Tabla 4.35. Precision media por Ventana Contextual

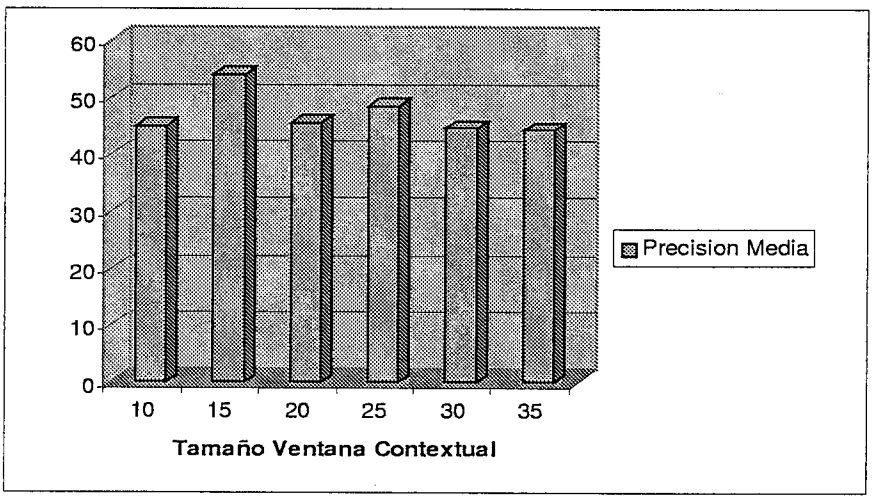

Figura 4.2. Mejor *precision* para las ventanas contextuales entre 10 y 35

Discusión. Se puede observar en las distintas tablas o gráficos presentados en este experimento que cuando la ventana contextual es de 15 palabras se obtiene la mayor "precisión" de nuestro método. Adicionalmente, y basándonos en el estudio, se puede de-

cir que una ventana de contexto que oscila entre 10 y 15 palabras obtendría los mejores resultados al aplicar el método de Marcas de Especificidad sin heurísticas. Por otro lado, y sobre la base de los resultados obtenidos, se puede concluir indicando que la venta óptima para el método propuesto está más cercana al tamaño de una frase que a la de un párrafo. Por ello, se decidió como ventana de contexto utilizar la frase, aunque no tiene un tamaño fijo, si que nos da indicios de contextualidad coherente en el proceso de desambiguación léxica[3].

4.3.2 Comparación de resultados

Experimento 4: Comparación con otros métodos.

Objetivo. Según Resnik y Yarowsky (1997), los métodos de WSD son difíciles de comparar unos contra otros, debido a la gran cantidad de diferencias que se encuentran en las palabras elegidas para desambiguar, además de los diferentes tipos de métodos empleados (métodos basados en el conocimiento, no supervisados, supervisados, en el corpus, etc). Sobre la base de las dificultades que nos podemos encontrar en la comparación de métodos, el objetivo de este experimento es realizar una comparación *directa* e *indirecta* del método de Marcas de Especificidad con otros métodos de WSD. Por comparación *directa* se entiende analizar los resultados obtenidos, cuando se aplican dos métodos sobre las mismas condiciones de experimentación para desambiguar. Sin embargo, comparación *indirecta* indica que se analizan los resultados de WSD con respecto a una tarea de Traducción Automática, Recuperación de Información, etc. Este experimento se centra en realizar una comparación *directa* con el método de Densidad Conceptual (Agirre y Rigau, 1996), ya que se dispone de la implementación de este método y podemos compararlo con los mismos datos. Y una comparación *indirecta* con el método de Yarowsky (1992)[4]

[3] Hay que indicar que las pruebas realizadas muestran que el tamaño medio oscila entre 10-15 palabras según el corpus.

[4] En (Agirre y Rigau, 1996) se compara el método de Densidad Conceptual con el de (Yarowsky, 1992), adaptándolo a los ficheros semánticoa de WordNet, en vez de usar las categorías semánticas del Roget's Thesaurus.

y Sussna (1993). Los datos de los tres últimos métodos se han obtenido de la comparación realizada en Agirre y Rigau (1996).

Descripción del recurso.

Comparación directa. Para la comparación *directa* del método de Marcas de Especificidad con el método de Densidad Conceptual se han utilizado las mismas palabras de los textos del corpus *Semantic Concordance (Semcor)* y un subconjunto de la enciclopedia electrónica *Microsoft Encarta 98 Encyclopedia Deluxe*, es decir, 624 nombres para el *Semcor* y 685 nombres para la enciclopedia *Encarta 98.*

Comparación indirecta. Para la comparación *indirecta* del método de Marcas de Especificidad con los métodos de Yarowsky (1992), Agirre y Rigau (1996) y Sussna (1993) se han utilizado diferentes corpus de entrada. Sin embargo, para el método de Marcas de Especificidad se han utilizado los mismos datos que en la comparación *directa*, es decir, las mismas palabras de los textos del corpus *Semantic Concordance (Semcor)* y de un subconjunto de la enciclopedia electrónica *Microsoft Encarta 98 Encyclopedia Deluxe.*

Tamaño del recurso.

Comparación directa. Los textos a desambiguar han sido escogidos al azar teniendo en cuenta que la ventana contextual es de una frase. El total de texto escogido es de 100 frases y 624 nombres para el Semcor y 100 frases y 685 nombres para la enciclopedia Encarta 98. Estas palabras serán las que sirvan de entrada a los métodos comentados anteriormente para hacer la comparación *directa.* Hay que puntualizar que la desambiguación de las palabras se hará sobre WordNet versión 1.6.

Comparación indirecta. Los textos elegido para el método de Marcas de Especificidad es el mismo que se ha utilizado para realizar la comparación *directa* y la desambiguación de este sistema será sobre WordNet versión 1.6. Sin embargo, para los otros tres métodos se han escogido 9000 palabras de un corpus de dominio público y la desambiguación será sobre WordNet versión 1.4.

Resultados.

Comparación directa. Los resultados que se mostrarán en este apartado son los obtenidos cuando se aplica el método de Marcas de Especificidad sin heurísticas y con heurísticas. Así, los resultados obtenidos al aplicar el conjunto de nombres elegidos de los corpus Semcor (624 nombres) y Encarta 98 (685 nombres) al método de Marcas de Especificidad sin heurísticas y al método de Densidad Conceptual se muestran en la Tabla 4.36.

SemCor					
Correctas		Incorrectas		No desambiguadas	
ME	DC	ME	DC	ME	DC
330	299	173	186	121	139
52,9%	47,84%	27,70%	29,76%	19,39%	22,40%
Encarta 98					
Correctas		Incorrectas		No desambiguadas	
ME	DC	ME	DC	ME	DC
383	337	181	233	121	115
55,9%	49,2%	26,40%	34,00%	17,60%	16,8%

Tabla 4.36. Resultados de Marcas de Especificidad sin heurísticas y Densidad Conceptual para SemCor y Encarta 98

Cuando se aplica el método de Marcas de Especificidad con heurísticas y el método de Densidad Conceptual al mismo conjunto de nombres anterior, se obtienen los resultados mostrados en la Tabla 4.37.

SemCor					
Correctas		Incorrectas		No desambiguadas	
ME	DC	ME	DC	ME	DC
415	299	200	186	9	139
66,4%	47,84%	32,00%	29,76%	1,50%	22,40%
Encarta 98					
Correctas		Incorrectas		No desambiguadas	
ME	DC	ME	DC	ME	DC
446	337	231	233	8	115
65,1%	49,2%	33,80%	34,00%	1,10%	16,8%

Tabla 4.37. Resultados de Marcas de Especificidad con heurísticas y Densidad Conceptual para SemCor y Encarta 98

Las dos Tablas 4.36, 4.37 muestran los porcentajes y los resultados de palabras correctas, incorrectas y no desambiguadas de cada uno de los dos métodos a comparar, método de Marcas de Especificidad (ME) con heurísticas y sin heurísticas y método de Densidad Conceptual (DC). Como comentario a los datos de la Tabla 4.37 y en particular a los presentados por el método de Marcas de Especificidad con heurísticas, hay que resaltar que los porcentajes de las palabras que quedan sin desambiguar, entre 1,1% y 1,5%, se deben a que no se obtiene ninguna marca de especificidad común que relacione a las palabras a ser desambiguadas. En cuanto a los porcentajes de las palabras que quedan sin desambiguar del Método de Densidad Conceptual, entre 16,8% y 22,4%, se deben a que el método devuelve varios sentidos posibles para una palabra. En este caso los resultados para el método de Marcas de Especificidad son bastante superiores a los de Densidad Conceptual. Sin embargo, una comparación más justa es la efectuada en la tabla 4.36, ya que se aplica el método de Marcas de Especificidad sin heurísticas, y en este caso se puede comprobar que los resultados son más parejos.

A continuación, se muestran los resultados de "cobertura absoluta", "precisión" y "cobertura" obtenidos al realizar la comparación *directa* en la Tabla 4.38.

SemCor			
%	**Cobertura absoluta**	**Precisión**	**Cobertura**
ME sin Heurísticas	80,60 %	65,60 %	52,90 %
ME con Heurísticas	98,56%	67,47 %	66,50 %
Densidad Conceptual	77,72 %	61,64 %	47,92 %
Encarta 98			
%	**Cobertura absoluta**	**Precisión**	**Cobertura**
ME con Heurísticas	82,30 %	67,90 %	55,90 %
ME con Heurísticas	98,83%	65,87%	65,10%
Densidad Conceptual	83,21%	59,12%	49,19%

Tabla 4.38. Resultados de cobertura absoluta, precisión y cobertura

Comparación indirecta. Los resultados obtenidos, al aplicar el conjunto de palabras elegidas de los corpus *SemCor* (624 pala-

bras) y *Encarta 98* (685 palabras) al método de Marcas de Especificidad y 9000 palabras de un corpus de dominio público al método de Densidad Conceptual y de Yarowsky, se muestran en la Tabla 4.39. Como sistema baseline se ha tenido en cuenta el sentido más frecuente de WordNet en el conjunto de palabras elegido de *SemCor* y *Encarta 98*.

%	Cobertura absoluta	Precisión	Cobertura
Marcas Especificidad	98,5%	66,6%	65,7%
Densidad Conceptual	86,2%	71,2%	61,4%
Yarowsky	100,0%	64,0%	64,0%
Sussna	100,0%	64,5%	64,5%
Baseline	100,0%	65,2%	65,2%

Tabla 4.39. Resultados de la comparación indirecta

Hay que aclarar que para el método de Marcas Especificidad se ha utilizado una ventana contextual de una frase, mientras que para los métodos de Yarowsky y Densidad Conceptual se ha utilizado una ventana contextual de 50 palabras y finalmente para el método de Sussna de 30 palabras. En definitiva, aunque se utilizan diferentes tamaños de ventanas contextuales y diferentes contextos de entrada, esta comparativa muestra que el método de Marcas de Especificidad funciona con una "cobertura" superior a los otros métodos y una "precisión" similar a los otros métodos.

Discusión.

Comparación directa. Una primera conclusión a destacar de los resultados obtenidos al aplicar los dos métodos a diferentes corpus es que, los dos métodos obtienen resultados muy similares sobre dos corpus con diferentes dominios. Esto demuestra que estos dos métodos pueden aplicarse a cualquier corpus con diferentes dominios y sus resultados seguirían siendo muy similares, por lo que se pueden considerar dos métodos muy estables.

Una segunda conclusión a destacar es que si se miran los resultados mostrados en la Tabla 4.37 el método de Densidad Conceptual deja muchas palabras sin desambiguar (139 y 115) con respecto al de Marcas de Especificidad (9 y 8). Esto es debido

a que el primero obtiene el mismo valor de densidad conceptual para diferentes sentidos y no es capaz de asignar un único sentido a la palabra que se desea desambiguar. Sin embargo, el método de Marcas de Especificidad en estos casos aplica la heurística de Marcas de Especificidad Común. Como se puede apreciar al aplicar esta heurística descienden considerablemente las palabras no desambiguadas, eso sí, no se sabe si esos sentidos asignados son correctos o incorrectos. Aunque observando los resultados de este experimento se puede afirmar que se obtienen más sentidos correctos que incorrectos. Si se observan los resultados de esta tabla, ME supera en un porcentaje bastante notable (un 66,4% y 65,1% frente a 47,89% y 49,2% para SemCor y Encarta 98, respectivamente) a DC en cuanto a los sentidos correctos asignados a las palabras. Evidentemente esto es debido a que DC deja muchas palabras sin desambiguar, ya que en cuanto a palabras desambiguadas incorrectamente los resultados son muy parecidos (32,00% y 33,80% frente a 29,76% y 34,00%).

Como último comentario, hay que decir que la Tabla 4.38 demuestra que los dos métodos tienen casi la misma "precisión" (67,47% y 65,87% frente a 61,64% y 59,52%), aunque la del método de Marcas de Especificidad es mejor en un 2%. Sin embargo, los resultados de "cobertura" para el método de Marcas de Especificidad superan a los de Densidad Conceptual en un porcentaje bastante considerable, aproximadamente de un 19% para *SemCor* y de un 16% para *Encarta 98*. Esto es debido a la decisión de tomar como no desambiguadas aquellos casos en que el método de Densidad Conceptual devuelve como salida varios sentidos para la misma palabra, por eso su valor de "cobertura absoluta" es bajo y afecta al valor de "cobertura" de dicho método. Si en este experimento se hubiera tomado la decisión de elegir el sentido más frecuente de WordNet, cuando el método de Densidad Conceptual devuelve varios sentidos para una misma palabra, los dos métodos tendrían valores muy similares tanto de "cobertura" como de "cobertura absoluta". Con este experimento se puede deducir que ambos métodos obtienen unos resultados de "precisión" y "cobertura" muy similares, aunque en este caso el valor

del método de Marcas de Especificidad haya sido mejor para la desambiguación del conjunto de palabras elegido.

Comparación indirecta. Las comparaciones que se han efectuado a los diferentes métodos se han realizado sobre diferentes versiones de WordNet y oraciones de entrada. Y por tanto, no pueden considerarse más que comparaciones indirectas. Por un lado, WordNet 1.6. es más polisémico que WordNet 1.4. y por otro, los corpus son de tamaños y características distintas. Teniendo en cuenta esto, el método de Marcas de Especificidad supera en "cobertura" y "precisión" al resto de métodos, menos en la "precisión" del método de Densidad Conceptual. Pero este resultado se puede justificar debido a que al tener una "cobertura absoluta" más baja su "precisión" tiene que aumentar, es decir se afina más al seleccionar los sentidos de las palabras pero a cambio desambigua menor número de palabras. Así, se puede decir que el método de Densidad Conceptual es más conservador que el resto.

Experimento 5: Comparación del método de Marcas de Especificidad con un modelo probabilístico basado en el principio de Máxima Entropía.

Objetivo. El principal objetivo de este experimento es evaluar y comparar el método de Marcas de Especificidad, el cual está basado en el conocimiento no supervisado y un método supervisado basado en los modelos de probabilidad de Máxima Entropía (MEX), descrito en (Suárez y Montoyo, 2001), sobre un mismo conjunto de ejemplos. Posteriormente a partir de la comparación de los resultados obtenidos se estudió la viabilidad de un nuevo método híbrido basado en la cooperación de ambos, es decir de un método basado en el conocimiento (Marcas de Especificidad) y de otro como un método de aprendizaje supervisado basado en el corpus (Máxima Entropia). Mediante la experimentación sobre un mismo conjunto de evaluación, el trabajo presentado muestra una mejora de un 12% cuando se combinan los dos métodos utilizados.

Descripción del recurso. Para evaluar y comparar los dos métodos aquí propuestos se han utilizado todos los artículos pre-

sentes en las carpetas Brown1 y Brown2 de *Semcor*, tal y como se distribuye con WordNet versión 1.6.

Tamaño del recurso. Para realizar el proceso de desambiguación se han seleccionado los siguientes nombres: *account, age, art, car, child, church, cost, duty, head, interest, line, member, people, term, test,* y *work.* En la realización de este experimento, se seleccionaron todas las oraciones en las que apareciera alguno de los nombres anteriormente seleccionados de entre todo el corpus *Semcor.* Para cada una de estas oraciones se forma el contexto de entrada mediante los nombres que acompañan a la palabra a desambiguar.

Resultados.

Marcas de Especificidad. Al aplicar el método de Marcas de Especificidad al contexto anteriormente comentado, este devuelve automáticamente el sentido correspondiente de WordNet para cada uno de los nombres. Los resultados[5] obtenidos cuando se aplica el método de Marcas de Especificidad sobre los nombres seleccionados se muestran en la Tabla 4.40.

Máxima Entropia. Para la construcción de los distintos clasificadores se han definido funciones que indican la presencia de palabras en posiciones relativas a la palabra objetivo: w-1, w-2, w-3, w+1, w+2, w+3, (w-1 w-2), (w-1, w+1), (w+1, w+2), (w-3, w-2, w-1), (w-2, w-1, w+1), (w-1 w+1 w+2), (w+1 w+2 w+3). También se han tenido en cuenta las etiquetas sintácticas en el contexto, nuevamente en posiciones relativas: p-3, p-2, p-1, p+1, p+2, p+3. Asimismo, se ha utilizado una función basada en la frecuencia de aparición de ciertas palabras con contenido léxico en contextos asociados a una clase en concreto: palabras que aparecen 5 veces[6] o más en todo el corpus de entrenamiento y para cada sentido de la palabra a clasificar. Para la evaluación del método

[5] A las medidas vistas anteriormente de "cobertura" (C), "precisión" (P) y "cobertura absoluta" (CA) se le añaden nuevos valores que se representará mediante el símbolo "#oc" y que mide la cantidad de ocurrencias de la palabra en el corpus evaluado, "#s" que mide el número de sentidos y "SMF" que es el sentido más frecuente en el corpus.

[6] Se evaluaron 3, 5 y 10 veces en el DSO, y el que mejor resultados dió fue el de 5 veces.

nombre	#oc	#s	SMF	P	C	CA
account	27	5	44,4%	48,0%	48,0%	100%
age	104	3	72,5%	52,3%	52,3%	100%
art	74	4	49,0%	33,3%	32,8%	98,4%
car	71	2	92,6%	73,4%	72,3%	98,5%
child	206	2	70,9%	62,2%	59,4%	95,6%
church	128	3	46,9%	53,9%	51,4%	95,3%
cost	85	3	88,2%	28,9%	28,9%	100%
duty	25	3	48,0%	34,8%	34,8%	100%
head	179	8	78,2%	20,4%	19,0%	93,5%
interest	139	7	41,0%	44,4%	44,4%	100%
line	124	22	16,1%	20,9%	20,3%	97,5%
member	74	3	90,5%	51,5%	51,5%	100%
people	281	4	89,7%	53,1%	52,0%	98,0%
term	55	5	67,3%	15,6%	15,6%	100%
test	36	6	52,8%	8,8%	8,8%	100%
work	208	6	41,3%	25,5%	25,3%	98,9%
TOTAL	**1810**	**5,37**	**60,4%**	**40,4%**	**39,5%**	**97,8%**

Tabla 4.40. Resultados de evaluación del método Marcas de Especificidad en *SemCor*

se ha dividido el corpus en 10 partes (*10-fold cross-validation*). Se realizan 10 experimentos distintos, seleccionando, en cada uno de ellos, una décima parte del corpus como conjunto de evaluación y el resto como conjunto de entrenamiento. Los resultados obtenidos se muestran en la Tabla 4.41. En ella se muestra la cantidad de ocurrencias, de sentidos de cada nombre en los conjuntos de evaluación y los valores promedio de "precisión", "cobertura" y "cobertura absoluta".

Cooperación de los dos métodos. Para comparar los dos métodos se ha revisado una porción del conjunto total de datos evaluados anteriormente. Esta porción se compone de 18 artículos del *Semcor* que contienen 267 ocurrencias de los nombres seleccionados. A partir de las evaluaciones de cada método, se han contabilizado aquellos casos en los que los dos métodos devuelven el mismo sentido para la misma ocurrencia. También se han contabilizado aquellos casos en que al menos uno de los métodos proporciona

nombre	#oc	#s	SMF	P	C	CA
account	27	5	44,4%	28,5%	26,3%	87,2%
age	104	3	72,5%	31,3%	14,3%	43,8%
art	74	4	49,0%	59,6%	57,5%	96,6%
car	71	2	92,6%	95,9%	95,9%	100%
child	206	2	70,9%	95,7%	16,9%	18,9%
church	128	3	46,9%	55,8%	54,3%	96,7%
cost	85	3	88,2%	88,3%	85,1%	96,2%
duty	25	3	48,0%	77,8%	68,5%	87,0%
head	179	8	78,2%	60,0	58,2%	96,1%
interest	139	7	41,0%	48,5%	45,4%	93,2%
line	124	22	16,1%	7,0%	6,7%	94,6%
member	74	3	90,5%	87,4%	87,4%	100%
people	281	4	89,7%	62,6%	35,9%	53,0%
term	55	5	67,3%	44,5%	43,0%	95,1%
test	36	6	52,8%	25,8%	25,2%	93,8%
work	208	6	41,3%	40,5%	39,2%	96,2%
TOTAL	**1810**	5,37	**60,4%**	**58,6%**	**47,3%**	**80,5%**

Tabla 4.41. Resultados de evaluación del método Máxima Entropia en *SemCor*

el sentido correcto. Un resumen de los resultados obtenidos se muestra en la Tabla 4.42.

Casos comparados	267
Coinciden	30,71%
Al menos uno acierta	71,91%
Aciertan los dos	22,47%
Fallan los dos	28,09%

Tabla 4.42. Comparativa de la combinación de resultados de los dos métodos

Discusión.

Marcas de Especificidad. Una ventaja importante al aplicar este método es su proceso no supervisado, por lo que no necesita procesos de entrenamiento, codificación manual léxica de las entradas ni etiquetado manual de los nombres del texto.

Máxima Entropia. Debido al tamaño del corpus *Semcor*, el método no dispone de una gran cantidad de ejemplos de los que aprender,

y la inmediata consecuencia de este hecho son los bajos valores de "cobertura" y "cobertura absoluta". Aunque el sistema podría utilizar alguna heurística (por ejemplo, frecuencia de aparición de cada clase en el corpus de entrenamiento) se asume que ante situaciones de información insuficiente la clasificación no se puede llevar a cabo. Asimismo, los valores tan dispares de "precisión" pueden ser imputables a la misma causa. El nombre *interest*, que en los resultados mostrados obtiene un 48,5% de "precisión" y un 45,4% de "cobertura", evaluado sobre el *DSO English sense-tagged corpus* (Martínez y Agirre, 2000), mucho más extenso, se llega a obtener un 74,5% y 73,8% respectivamente. Debido a diferencias en el etiquetado entre *DSO* y *Semcor* no es posible aplicar el aprendizaje con el primero al segundo.

Cooperación de los dos métodos. A la vista de los datos mostrados en la Tabla 4.42, se puede observar que la cooperación de los dos métodos mejora la desambiguación del sentido de las palabras, ya que los porcentajes de "precisión", y "cobertura" son mejores que los obtenidos por cada método individualmente. En concreto, para la porción del *Semcor* estudiada, la cooperación entre los dos podría mejorar en aproximadamente un 12% los resultados del método con más éxito, ya que alguno de los dos métodos (o los dos) obtiene la desambiguación correcta en el 72% de los casos. Ambos métodos coinciden y aciertan en el 22,5% de los casos. Al respecto de esto último, es remarcable el dato de que en los casos de coincidencia la mayor parte corresponde a aciertos: del 31% de los casos en que ambos proponen el mismo sentido, los aciertos suponen el 73%.

Los resultados de los experimentos realizados demuestran que la cooperación de ambos métodos mejora la desambiguación del sentido de las palabras si los comparamos con los obtenidos de sus respectivas aplicaciones individuales. Con estos datos se puede afirmar que un nuevo método híbrido que combine las aproximaciones basadas en el conocimiento y en el corpus mejorará la desambiguación del sentido de las palabras.

Los métodos basados en corpus tienen la desventaja de su dependencia del propio corpus de entrenamiento, tanto por disponibilidad e idoneidad del mismo como por su discutible aplicación

en diferentes dominios (Escudero et al., 2000a). Los sistemas basados en el conocimiento obtienen, en el momento actual, peores resultados que los primeros y exigen un gran esfuerzo de representación del conocimiento, pero son más fácilmente aplicables a dominios heterogéneos. Un método que una las dos aproximaciones aprovecharía las ventajas de cada una y minimizaría los aspectos negativos respectivos.

4.3.3 Evaluación final

Experimento 6: Evaluación en Senseval-2. Hoy en día, hay muchos sistemas que determinan automáticamente el sentido de las palabras en un contexto determinado. Pero sin embargo, es muy difícil realizar una comparación directa de cada uno de ellos para ver cual es el que mejor resuelve el problema expuesto. SENSEVAL[7] trata de resolver este problema.

La última competición, denominada SENSEVAL-2 (*Second International Workshop on Evaluating Word Sense Disambiguation Systems*) formó parte del evento ACL SIGLEX y se celebró desde el 1 Noviembre del 2000 hasta el 5/6 de Julio de 2001 que tuvo lugar el Workshop en Toulouse (France). Para esta competición se definieron 3 tipos de tareas sobre 12 lenguajes distintos. A continuación se describen cada una de las tareas y sobre los lenguas que intervienen.

- La tarea *"all-words"* consiste en encontrar el sentido a todas las palabras de los textos seleccionados por el comité y evaluar todas esas palabras. Esta tarea se realizó a los cuatro lenguas siguientes: Checo, Holandés, Inglés y Estonio.
- En la tarea *"lexical sample"*, en primer lugar se eligen un conjunto de palabras de un lexicón, posteriormente se deben encontrar los sentidos de esas palabras a partir de los ejemplos de contexto seleccionados por el comité y finalmente se evalúan solamente esas palabras. Esta tarea se realizó a los siguientes

[7] SENSEVAL es una competición que propone una colección de evaluación, un conjunto de definiciones y una serie de métricas comunes, y que pretende evaluar la potencia o debilidad de los sistemas WSD existentes en diferentes lenguas. Puede encontrarse más información en: `http://www.sle.sharp.co.uk/senseval2/`.

nueve lenguajes: Vasco, Chino, Danés, Inglés, Italiano, Japonés, Coreano, Español, Sueco.

- La tarea *"translation task"*, consiste en encontrar el sentido correspondiente de una palabra dada para las distintas traducciones a otras lenguas. Esta tarea se realizó sólo para el Japonés.

Una información más detallada sobre los datos de entrenamiento, prueba y resultados de la evaluación SENSEVAL-2 se pueden encontrar en `http://www.sle.sharp.co.uk/senseval2/`.

El sistema denominado *"The University of Alicante Word Sense Disambiguation System"*, el cual está basado en el método de Marcas de Especificidad (para desambiguar nombres) y en los modelos probabilísticos de Máxima Entropía (Suárez y Palomar, 2002) (para verbos y adjetivos), participó en la competición del SENSEVAL-2. Este sistema participó en la tarea de *"lexical sample"* para el inglés y español.

Objetivo. Después de presentar brevemente en que consiste la competición SENSEVAL-2, sus tareas y qué lenguas intervienen en cada una de ellas, el objetivo del presente experimento es mostrar los resultados obtenidos por el método de Marcas de Especificidad para los nombres seleccionados por el comité en la tarea de "lexical sample" tanto para inglés como para español.

Descripción del recurso.

Para Inglés. Los corpus en inglés para la competición SENSEVAL-2 se obtienen a partir de los recursos BNC-2 y Penn Treebank (compuesto a partir de *Wall Street Journal, Brown y IBM manuals*). Ambos recursos están disponibles públicamente.

Los ficheros del corpus suministrados para el proceso de entrenamiento y de prueba tienen formato XML. Además, todas las ocurrencias de entrada, correspondientes a la palabra a desambiguar en la tarea *"lexical sample"*, vienen etiquetados con la categoría sintáctica de la palabra (nombre, verbo o adjetivo).

El diccionario utilizado para proporcionar los sentidos a las palabras es el WordNet 1.7 (versión pública en Abril de 2001), el cual ha sido revisado y actualizado y sigue manteniendo el mismo formato que WordNet 1.6. especificado por Princeton. Hay que

comentar que el método de Marcas de Especificidad se adaptó a la versión 1.7 de WordNet, ya que los ejemplos de "test" también estaban marcados con esta versión. Además, para esta prueba se usó de contexto a todo el texto proporcionado por cada una de los ejemplos en vez de frase a frase.

La tarea *"lexical sample"* para el inglés se ha realizado sobre 71 palabras, de las cuales 29 eran nombres, 28 verbos y 14 adjetivos. Los resultados obtenidos para nombres se muestran en la Tabla 4.43, en donde se muestra la categoría léxica de las palabras (POS), el número de ejemplos (incluyendo datos del "test" y "train"), el número de sentidos y los resultados obtenidos por nuestro sistema (contando el número aciertos).

Los resultados obtenidos para adjetivos se muestran en la Tabla 4.44, en donde se muestra la categoría léxica de las palabras (POS), el número de ejemplos (incluyendo datos del "test" y "train"), el número de sentidos y los resultados obtenidos por nuestro sistema.

Los resultados obtenidos para verbos se muestran en la Tabla 4.45, en donde se muestra la categoría léxica de las palabras (POS), el número de ejemplos (incluyendo sólo datos del "test") y los resultados obtenidos por nuestro sistema.

Para Español. Los corpus en español para la competición SENSEVAL-2 se obtienen a partir de dos recursos: El Periódico (un periódico español) y Lexesp-III (una colección de textos de diferentes dominios; DGICYT APC 99-0105). Los ficheros del corpus suministrados para el proceso de entrenamiento y de test tienen formato XML

El diccionario utilizado ha sido creado específicamente para esta tarea y consiste de una definición para cada sentido, unido al numero de sentido correspondiente a esa palabra, la categoría sintáctica y algunas veces a ejemplos y sinónimos. No han sido consideradas las palabras compuestas ni los nombres propios.

La tarea *"lexical sample"* para el Español se ha realizado sobre 40 palabras, de las cuales 18 eran nombres, 13 verbos y 9 adjetivos. Los resultados obtenidos para nombres se muestran en la Tabla 4.46, para los verbos en la Tabla 4.47 y para los adjetivos en la Tabla 4.48. Además, para cada una de las tablas se muestran la

Palabra	Pos	# Ejemplos	# Sentidos	Sistema
Art	n	294	5	34,7%
Authority	n	276	7	15,22%
Bar	n	455	13	25,52%
Bum	n	137	4	27,27%
Chair	n	207	4	14,70%
Channel	n	218	7	20,55%
Child	n	193	4	35,93%
Church	n	192	3	39,06%
Circuit	n	255	6	27,06%
Day	n	434	9	5,5%
Detention	n	95	2	25,0%
Dyke	n	86	2	85,7%
Facility	n	172	5	20,68%
Fatigue	n	128	4	9,3%
Feeling	n	153	6	27,45%
Grip	n	153	7	21,57%
Hearth	n	96	3	28,12%
Holiday	n	93	2	21,61%
Lady	n	158	3	50,94%
Material	n	209	5	31,88%
Mouth	n	179	8	1,6%
Nation	n	112	3	16,21%
Nature	n	138	5	32,61%
Post	n	236	8	10,12%
Restraint	n	136	6	22,72%
Sense	n	160	5	26,41%
Spade	n	98	3	69,69%
Stress	n	118	5	10,81%
Yew	n	85	2	32,14%
Total		5266	5	27,24%

Tabla 4.43. Resultados de nombres para la tarea *"lexical sample"* de inglés

categoría léxica de las palabras (POS), el número de ejemplos (incluyendo datos del "test" y "train"), el número de sentidos y el porcentaje del sentido más frecuente (SMF).

Las ocurrencias (incluyendo datos de "test" y "train") elegidas solamente pueden pertenecer a una categoría sintáctica, así se eligieron 2336 ocurrencias a desambiguar para nombres, 2276

Palabra	Pos	# Ejemplos	# Sentidos	Sistema
Blind	a	163	3	81,81%
Colorless	a	103	2	0%
Cool	a	158	6	42,31%
Faithful	a	70	3	73,91%
Fine	a	212	9	52,86%
Fit	a	86	3	24,14%
Free	a	247	8	54,88%
Graceful	a	85	2	79,31%
Green	a	284	7	69,15%
Local	a	113	3	52,63%
Natural	a	309	10	0%
Oblique	a	86	2	75,86%
Simple	a	196	7	56,06%
Solemn	a	77	2	96%
Vital	a	112	4	94,7%
Total		2301	4,73	56,9%

Tabla 4.44. Resultados de adjetivos para la tarea *"lexical sample"* de inglés

para verbos y 2093 para adjetivos. En total, se pasaron 6705 ocurrencias para la tarea de *"lexical sample"* de español sobre las palabras anteriormente mostradas.

Resultados. A continuación se mostrarán los resultados obtenidos en la competición SENSEVAL-2 para el inglés y español por el sistema denominado *"The University of Alicante Word Sense Disambiguation System"*. En primer lugar, se mostrarán los resultados oficiales globales del sistema presentados por el comité del SENSEVAL-2 y a continuación se presentarán los resultados desglosados para nombres, verbos y adjetivos. Así, se conocerá la efectividad de los dos métodos que componen al sistema global.

Para Inglés. Los resultados oficiales provisionales tal cual fueron enviados por el comité del SENSEVAL-2 son los siguientes:

```
Fine-grained score for "montoyo-Univ._Alicante_System":
 precision: 0.421 (1779.20 correct of 4230.90 attempted)
 recall: 0.411 (1779.20 correct of 4328.00 in total)
 attempted: 97.756 % (4230.90 attempted of 4328.00 in total)

Coarse-grained score for "montoyo-Univ._Alicante_System":
```

Palabra	POS	#Ejemplos	Sistema
Begin	v	280	69,64%
Call	v	66	36,36%
Carry	v	66	39,39%
Collabora	v	30	90,0%
Develop	v	69	36,23%
Draw	v	41	14,63%
Dress	v	59	54,24%
Drift	v	32	40,62%
Drive	v	42	35,71%
Face	v	93	78,49%
Find	v	68	17,64%
Keep	v	67	35,82%
Leave	v	66	37,87%
Live	v	67	55,22%
Match	v	42	64,28%
Play	v	66	46,97%
Pull	v	60	15%
Replace	v	45	46,66%
See	v	69	37,68%
Serve	v	51	49,02%
Strike	v	54	27,77%
Train	v	63	42,85%
Treat	v	44	52,27%
Turn	v	67	28,36%
Use	v	76	67,10%
Wander	v	50	74,0%
Wash	v	12	33,33%
Work	v	60	28,33%
Total		1805	44,83%

Tabla 4.45. Resultados de verbos para la tarea *"lexical sample"* de inglés

```
precision: 0.526 (2226.40 correct of 4230.90 attempted)
recall: 0.514 (2226.40 correct of 4328.00 in total)
attempted: 97.756 % (4230.90 attempted of 4328.00 in total)
```

Los resultados presentados anteriormente se dividen en dos métodos distintos de puntuar, los correspondientes a *"Fine-grained"* y a *"Coarse-grained"*. El método de *"Fine-grained"* consiste en contar como sentido correcto solo el primer sentido etiquetado

Palabra	Pos	# Ejemplos	# Sentidos	SMF	Cobertura
Autoridad	n	122	6	49%	68%
Bomba	n	113	2	71%	27%
Canal	n	156	5	33%	34%
Circuito	n	123	4	34%	43%
Corazón	n	146	5	36%	23%
Corona	n	119	4	45%	53%
Gracia	n	160	6	30%	28%
Grano	n	78	3	44%	37%
Hermano	n	135	5	61%	74%
Masa	n	131	5	45%	39%
Naturaleza	n	167	10	44%	45%
Operación	n	142	5	35%	71%
Organo	n	212	4	52%	73%
partido	n	159	2	55%	81%
pasaje	n	112	4	39%	83%
Programa	n	142	6	49%	36%
Tabla	n	119	3	51%	88%
Nombres		2336	4	45%	55%

Tabla 4.46. Resultados de Nombres para la tarea *"lexical sample"* de español

Palabra	Pos	# Ejemplos	# Sentidos	SMF	Cobertura
Actuar	v	155	6	28%	27%
Apoyar	v	210	4	64%	63%
Apuntar	v	191	8	47%	55%
Clavar	v	131	9	44%	50%
Conducir	v	150	9	35%	35%
Copiar	v	147	8	32%	42%
Coronar	v	244	6	32%	49%
Explotar	v	133	6	32%	49%
Saltar	v	137	14	15%	51%
Tocar	v	236	12	31%	51%
Tratar	v	192	13	21%	39%
Usar	v	167	4	68%	77%
Vencer	v	183	8	63%	72%
Verbos		2276	7	40%	51%

Tabla 4.47. Resultados de verbos para la tarea *"lexical sample"* de español

Palabra	Pos	# Ejemplos	# Sentidos	SMF	Cobertura
brillante	a	256	2	52%	63%
Ciego	a	114	4	54%	71%
Claro	a	204	7	83%	82%
Local	a	139	3	74%	84%
Natural	a	137	6	25%	34%
popular	a	661	3	65%	77%
simple	a	217	5	61%	67%
Verde	a	109	9	37%	48%
Vital	a	256	4	45%	65%
Adjetivos		2093	4	58%	66%

Tabla 4.48. Resultados de adjetivos para la tarea *"lexical sample"* de español

ignorando los restantes, ya que a una palabra el sistema le puede asignar más de un sentido etiquetado con una probabilidad. El método de *"Coarse-grained"* consiste en contar como sentido correcto aquellos casos en que esté etiquetado en el subconjunto de sentidos etiquetados por el sistema para la palabra a desambiguar. De acuerdo a la aplicación de estos métodos de puntuación, es evidente, y así lo confirman los resultados anteriores, que los resultados presentados por *"Coarse-grained"* son bastante mejores que los de *"Fine-grained"*.

Los resultados para la tarea de *"lexical sample"* en inglés se muestran en la Tabla 4.49, desglosando los resultados para nombres, verbos y adjetivos.

POS	Precisión	Cobertura
Nombres	30 %	29,24 %
Verbos	48,6 %	44,83 %
Adjetivos	70,9 %	56,9 %

Tabla 4.49. Resultados de *"lexical sample"* en inglés (Fine-grained)

En la Tabla 4.49 se pueden observar los resultados de "precisión" y "cobertura" obtenidos al desambiguar los nombres, verbos y adjetivos. Los nombres tienen un porcentaje más bajo porque en esta prueba sólo se evalúan unas palabras en concreto todas

ellas polisémicas. Mientras que en los otros experimentos sobre SemCor y Encarta se está trabajando con todos los nombres y no todos polisémicos, lo cual ayuda a afinar más en la jerarquía de WordNet y se obtiene una mejor desambiguación.

Además, el método de Marcas de Especificidad es un método no-supervisado basado en el conocimiento y si se comparara con los métodos participantes no-supervisados que han participado en esta tarea, este estaría mejor clasificado.

Para Español. Los resultados oficiales provisionales tal cual fueron enviados por el comité del SENSEVAL-2 son los siguientes:

```
Fine-grained score for "montoyo-Univ. Alicante System":
 precision: 0.514 (1118.20 correct of 2176.90 attempted)
 recall: 0.503 (1118.20 correct of 2225.00 in total)
 attempted: 97.838 % (2176.90 attempted of 2225.00 in total)
```

Como se puede observar, en la tarea de *"lexical sample"* en español solo se utilizó la política de *"Fine-grained"* para calcular la "precisión" y "cobertura" del sistema.

Los resultados para la tarea de *"lexical sample"* en español se muestran en la Tabla 4.50, con el desglose de los resultados para nombres, verbos y adjetivos.

POS	Precisión	Cobertura
Nombres	56,6 %	55 %
Verbos	51,1 %	51 %
Adjetivos	68,7 %	66 %

Tabla 4.50. Resultados de *"lexical sample"* en español (Fine-grained)

Discusión. Antes de empezar a discutir los resultados obtenidos por el sistema *"The University of Alicante Word Sense Disambiguation System"*, que tomó parte en la competición SENSEVAL-2 para la tarea de *"lexical sample"* en inglés y español, hay que recordar lo siguiente. Este sistema se clasifica en la competición dentro de los métodos supervisados, pero sin embargo uno de los métodos (método de Marcas de Especificidad) que actúan en este sistema para desambiguar los nombres está clasificado en los

métodos no-supervisados. Por lo tanto, esta discusión se centra en los resultados de la competición obtenidos sobre los nombres.

Para Inglés. Con el objetivo de centrar la discusión, en la Figura 4.3 se muestra una gráfica ilustrativa de los resultados obtenidos por todos los sistemas, tanto supervisados como no-supervisados, partícipes en la tarea de *"lexical sample"* para el inglés.

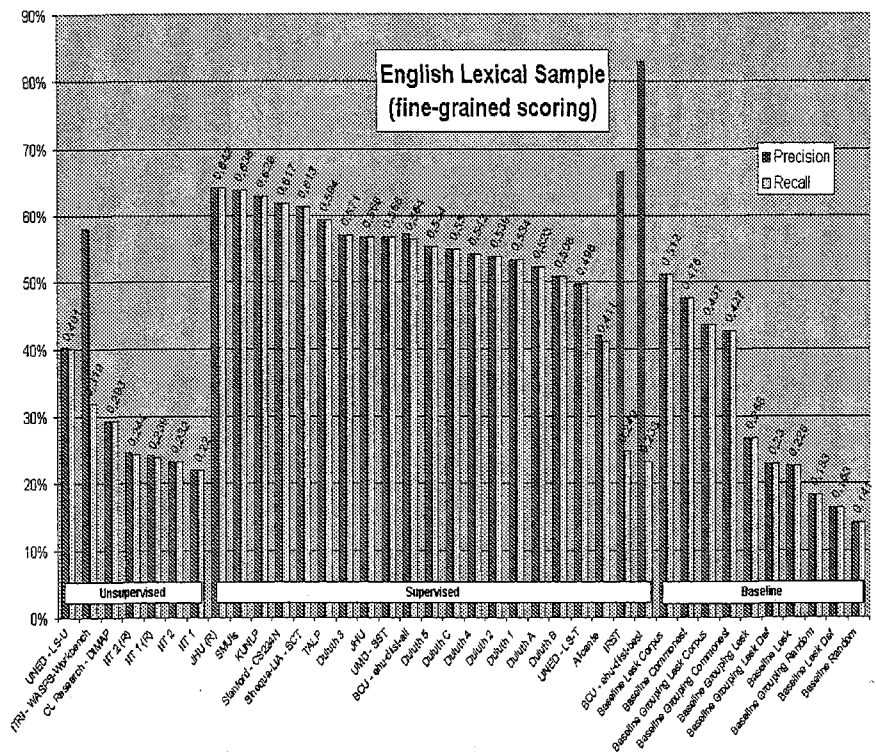

Figura 4.3. Resultados de todos los sistemas para *"lexical sample"* en inglés

Analizando los resultados de los nombres (30 % de "precisión" y 29,24 % de "cobertura") mostrados en la Tabla 4.49 de la sección anterior se observa que son más bajos que los obtenidos para verbos y adjetivos. Esto es debido al tipo de método usado para desambiguar los nombres, que utiliza un método basado en el conocimiento (no-supervisado), frente a los verbos y adjetivos que han utilizado un método basado en el corpus (supervisado).

Sin embargo, si se comparan estos valores con los obtenidos por los sistemas participantes en la competición clasificados como no-supervisados, estos serían de los mejores resultados. Se puede ver en la figura 4.3 que los dos mejores sistemas no-supervisados obtienen unos porcentajes entre 40 % y 31 %, y el resto ya obtienen porcentajes por debajo de 30 %, es decir, más pobres que el nuestro.

Un problema detectado al desambiguar los nombres fue que no tratamos las palabras compuestas. Si se hubieran tratado se hubieran obtenido mejores resultados, ya que, una palabra compuesta como *industrial plant* que forme parte del contexto de la palabra a desambiguar afina mucho más el sentido que si se trata, por una parte *industrial* y por otra *plant*. Otra conclusión importante es realizar un exhaustivo preproceso de las ocurrencias entregadas por el comité para cada palabra a desambiguar, ya que la información suministrada por el contexto se enriquecería con reconocimiento de entidades, análisis total, etc. Y este tipo de información beneficiaría al método de Marcas de Especificidad para obtener más relaciones entre conceptos de WordNet, etc.

Para Español. Igual que se hizo para el inglés, a continuación, mediante la figura 4.4, se muestra una gráfica ilustrativa de los resultados obtenidos por todos los sistemas supervisados partícipes en la tarea de *"lexical sample"* para el español.

El método de Marcas de Especificidad se aplicó directamente sobre el Spanish WordNet con todas las heurísticas, pero se tuvieron que adaptar las que utilizan las glosas de WordNet para realizar la desambiguación. Para ello se tradujeron las nombres en inglés a español de las glosas y a partir de ellas se aplicaron estas heurísticas.

Si se observan los resultados de los nombres (56,6 % de *precision* y 43,5 % de *recall*) mostrados en la Tabla 4.50 de la sección anterior, resalta a la vista que estos resultados son bastante mejores que los del inglés. Esto es debido a que el diccionario en español agrupaba varios sentidos de WordNet, por lo tanto como hacía un *clustering* previo, se obtenían mejores resultados. Por lo tanto el Sistema de Marcas de Especificidad se beneficia de esto, mejorando los resultados de desambiguación. Además,

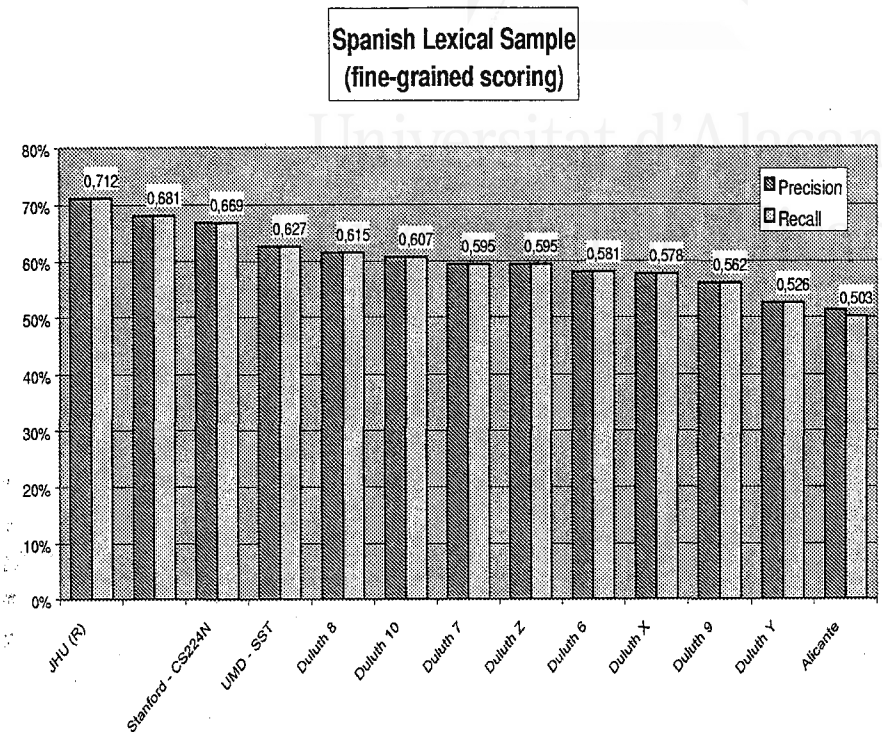

Figura 4.4. Resultados de todos los sistemas para *"lexical sample"* en español

en esta tarea se introdujeron algunas mejoras como por ejemplo analizar y tratar las definiciones de las palabras en el diccionario aportado por el comité para esta tarea. Una aclaración sobre la diferencia entre los datos de "precisión" (56,6 %) y de "cobertura" (43,5 %) es que hubo un error en la implementación de las palabras acentuadas (*corazón*, *operación* and *órgano*). Por tal motivo, todas las ocurrencias para estas tres palabras no pudieron desambiguarse, repercutiendo en el resultado de la "cobertura". Si estas tres palabras se hubieran desambiguado como el resto de palabras, el resultado de "precisión" hubiera aumentado en un porcentaje bastante menor al de "cobertura". Sin embargo, la ´´cobertura" se hubiera casi igualado con la "precisión", ya que este sistema está diseñado para que se deje pocas palabras sin desambiguar y por consiguiente tiende a igualar estas dos medidas.

Se pueden ver en la figura 4.4 que los cinco primeros sistemas supervisados obtienen unos porcentajes entre 71,2 % y 61,5 %, y el resto ya obtienen porcentajes muy similares entre 60,7 y 50,3 %. Con estos datos hay que resaltar que todos los sistemas son supervisados, aunque el sistema *"The University of Alicante Word Sense Disambiguation System"* tiene parte de no-supervisado. Por lo tanto, si se hubiera utilizado el método de Marcas de Especificidad (basado conocimiento y no-supervisado) para desambiguar también a los verbos y adjetivos, y tomando como premisa que el sistema desambigua en un 56,6 %, este sistema estaría en un nivel muy próximo a los sistemas intermedios de la clasificación. Por lo tanto, se puede afirmar que para ser un método no-supervisado está obteniendo unos resultados bastante satisfactorios.

Experimento 7: Evaluación final sobre todo el corpus *SemCor*.

Objetivo. El objetivo del presente experimento es mostrar los resultados de desambiguación obtenidos por el método de Marcas de Especificidad sobre todos los documentos de la colección de SemCor. Para llevar a cabo este trabajo se ha aplicado de distinta manera el método, con el objetivo de comprobar y verificar cuando y cómo el método de Marcas de Especificidad funciona mejor o peor, si cuando las heurísticas se aplican en cascada[8] en un orden preestablecido o cuando se aplican independientemente[9].

También se presenta en este experimento una comparación entre nuestro método y el de Densidad Conceptual (Agirre y Rigau, 1996), una variante del algoritmo de Densidad Conceptual propuesto por UNED (Fernández-Amorós et al., 2001a), el algoritmo de Lesk (Lesk, 1986) y la medida de la heurística del sentido más frecuente (coger siempre el primer sentido de WordNet) sobre los mismos datos de entrada, es decir sobre todo el SemCor.

Descripción del recurso. La evaluación ha sido realizada sobre la colección de documentos SemCor. Este consiste de un conjunto

[8] Aplicar las heurísticas en cascada se entiende que en primer lugar se aplica el método base y si hay palabras sin desambiguar se aplica la primera heurísticas y así sucesivamente hasta que se apliquen todas.

[9] Las heurísticas se aplican siempre que se puedan y se combine el resultado de todas ellas.

de 171 documentos donde todas las palabras son anotadas con el sentido más apropiado de WordNet. En esta evaluación, nuestro método como todos los comentados anteriormente se han ejecutado sobre todas las palabras cuya categoría gramatical es nombre correspondientes a todos los documentos del Semcor.

SemCor es un subconjunto del Brown Corpus, etiquetado con información semántica por el mismo equipo creador de WordNet. Todos los nombres, verbos, adjetivos y adverbios están etiquetados con el sentido correspondiente a la base de datos léxica WordNet. La construcción del SemCor se realizó de forma manual con la ayuda de herramientas software diseñadas especialmente para este propósito. SemCor puede ser utilizado para entrenar sistemas en la resolución de la ambigüedad léxica, pues se pueden extraer contextos (con sus palabras circundantes) de diferentes sentidos de palabras polisémicas.

Resultados. A continuación se mostrarán los resultados obtenidos por el método de Marcas de Especificidad cuando se aplica con las heurísticas en cascada o independientemente sobre todos los documentos de la colección SemCor.

Método con heurísticas aplicadas en cascada con ventana de frase y en el orden presentado. Los resultados obtenidos para este caso se muestran en la Tabla 4.51, teniendo en cuenta conjuntamente los nombres polisémicos y monosémicos .

Método WSD	Precisión	Cobertura	Cobertura Absoluta
Marcas de Especificidad	55,3%	52,2%	94,3%

Tabla 4.51. Resultados del método con heurísticas aplicadas en cascada con nombres polisémicos y monosémicos

Si se tienen en cuenta sólo las palabras polisémicas, se obtienen los resultados mostrados en la Tabla 4.52.

Método con heurísticas aplicadas independientemente. Los resultados[10] obtenidos para este caso se muestran en la Tabla 4.53,

[10] Las heurísticas de hipónimo y glosa de hipónimo se han aplicado sobre 10 ficheros de SemCor debido al tiempo de computación tan elevado.

Método WSD	Precisión	Cobertura	Cobertura Absoluta
Marcas de Especificidad	37,65%	31,11%	94,3%

Tabla 4.52. Resultados del método con heurísticas aplicadas en cascada con nombres polisémicos

teniendo en cuenta conjuntamente los nombres polisémicos y monosémicos.

Heurísticas	Precisión	Cobertura	Cobertura Absoluta
H. Hiperónimo	56,3 %	44,7 %	79,5 %
H. Definición	48,0 %	36,3 %	75,8 %
H. Hipónimo	55,6 %	43,6 %	78,4 %
H. Glosa Hiperónimo	55,5 %	45,0 %	81,1 %
H. Glosa Hipónimo	61,7 %	49,4 %	79,8 %
H. MEC	56,5 %	44,3 %	78,4 %

Tabla 4.53. Resultados del método con heurísticas aplicadas independientemente sobre nombres polisémicos y monosémicos

Si sólo se tienen en cuenta los nombres polisémicos, se obtienen los resultados que se muestran en la Tabla 4.54.

Heurísticas	Precisión	Cobertura	Cobertura Absoluta
H. Hiperónimo	42,0 %	31,3 %	74,5 %
H. Definición	30,0 %	20,9 %	69,9 %
H. Hipónimo	39,3 %	28,5 %	72,6 %
H. Glosa Hiperónimo.	41,2 %	31,6 %	76,4 %
H. Glosa Hipónimo.	48,1 %	35,8 %	74,5 %
H. MEC	42,3 %	31,0 %	73,2 %

Tabla 4.54. Resultados del método con heurísticas aplicadas independientemente sobre nombres polisémicos

A continuación se ha realizado una evaluación conjunta de todas estas heurísticas. La forma de realizar esta prueba es teniendo en cuenta que cada heurística está votando con un peso (en nuestro caso el peso es 1, aunque se podría normalizar entre 0 y 1) para

cada sentido y entonces el resultado obtenido para cada sentido se suma y se realiza una combinación o voting. Por ejemplo, si hay cuatro heurísticas que votan por el mismo sentido y las otras 2 por otro, entonces a la palabra a desambiguar se le asignará el sentido con una votación de 4 (la mayor votación). La evaluación utilizando voting se ha realizado sobre una colección de 10 ficheros de SemCor. Los resultados obtenidos son de un 54,6 % de cobertura para palabras polisémicas y monosémicas y de un 38,9 de cobertura para palabras polisémicas solamente. En cuanto a la precisión es de 56,7 % para palabras polisémicas y monosémicas y 42,6 % para palabras polisémicas.

Como se ha comentado en los objetivos, en este experimento se presenta una comparación entre nuestro método y el de Densidad Conceptual (Agirre y Rigau, 1996), una variante del algoritmo de Densidad Conceptual propuesto por UNED (Fernández-Amorós et al., 2001a), el algoritmo de Lesk (Lesk, 1986) y la heurítica del sentido más frecuente (coger siempre el primer sentido de WordNet) sobre los mismos datos de entrada, es decir sobre todo el SemCor. La Tabla 4.55 muestra los resultados de "cobertura" obtenidos al aplicar los distintos métodos a toda la colección de documentos del SemCor.

Método WSD	Cobertura
ME con Heurísticas	31,1%
Método UNED	31,3%
Lesk	27,4%
Densidad Conceptual	22,0%
H. más frecuente WordNet	70%

Tabla 4.55. Comparación de métodos WSD

Discusión. Nuestro método de Marcas de Especificidad obtiene un 31,1 % de "cobertura", el cual junto con el método de la UNED obtienen los mejores resultados de los métodos comparados. Sin embargo, todavía está muy lejos del 70 % que obtiene la heurística más frecuente. Esto nos hace pensar que todos los métodos mostrados anteriormente, los cuales se basan en las rela-

ciones conceptuales como información para desambiguar el sentido de las palabras, deberían ser rechazados para la tarea de WSD. Sin embargo, esto es una errónea conclusión, por las siguientes razones:

- Las anotaciones manuales tienden a elegir el primer sentido de WordNet (el cual es el sentido más frecuente). Por ejemplo, cuando en una anotación manual se quieren desambiguar todas las palabras, y la persona que está realizando esta tarea tiene que seleccionar el sentido apropiado de cada una de las ocurrencias, las cuales de media tiene 5 ó más sentidos diferentes. En estos casos la persona que está anotando las palabras tiende a seleccionar el primer sentido que parece acorde al contexto, y esto produce una tendencia a elegir los primeros sentidos del rango correspondiente. Por eso distintos estudios sobre cómo evaluar a los sistemas WSD (Resnik y Yarowsky, 1999; Kilgarriff y Rosenzweig, 2000) están a favor de la notación "lexical sample" que se utilizó en el SENSEVAL-1 y SENSEVAL-2, ya que el anotador selecciona el sentido de las ocurrencias de una misma palabra, lo cual permite familiarizarte con los sentidos de esa palabra.
- Además de los problemas presentados en la anotación manual, la tarea de anotar todas las palabras de un texto implica que el sistema debe repetitivamente intentar desambiguar ejemplos de términos muy comunes, los cuales pueden tener 20 sentidos diferentes y que son muy difíciles de desambiguar.
- Otra razón es que este método sólo tiene en cuenta nombres y muchas veces son los adjetivos o verbos del contexto los que nos permiten desambiguar.
- Además, en este método se trata básicamente información contenida en WordNet (limitada) y se le da mucha importancia a las relaciones clase/subclase.
- Por eso, este método en realidad funcionará muy bien para desambiguar palabras fuertemente relacionadas por relaciones clase/subclase como se verá a continuación en los siguientes experimentos del próximo capítulo. Pero no está pensado para desambiguar textos.

Yo creo que una conclusión acertada es que ni las medidas conceptuales ni las contextuales son suficientes por si solas para realizar una correcta tarea de WSD, y que necesitan de otras técnicas. Por eso este método no nos permite desambiguar completamente pero se puede proponer como una componente de un sistema más completo basado en el conocimiento o colaborar con uno supervisado, etc.

Volver al índice/Tornar a l'índex

5. Enriquecimiento de WordNet con Sistemas de Clasificación

En este capítulo se presenta la aplicación del método de Marcas de Especificidad para enriquecer semánticamente WordNet con etiquetas de dominio o categorías preestablecidas en otros sistemas de clasificación. El sistema de clasificación utilizado para etiquetar automáticamente los *synsets* de WordNet ha sido el *Subject Reference System* desarrollado por IPTC[1]

Este capítulo ha sido organizado de la siguiente forma. A continuación se presentan distintos sistemas de clasificación y en particular el sistema IPTC. En segundo lugar se describe detalladamente el método propuesto para enriquecer semánticamente WordNet con categorías IPTC (Versión IPTC/1). Posteriormente, se describen las características del diseño e implementación de la interfaz construida para extender y mejorar la base de datos léxica WordNet. Y finalmente, se muestran los experimentos realizados al método propuesto así como las conclusiones obtenidas al analizar sus resultados.

5.1 Introducción

Algunos investigadores han propuesto varias técnicas para tomar ventaja de la utilización de más de un recurso léxico, mediante la integración de varios recursos léxicos a partir de algunos preexistentes. Por ejemplo, el trabajo de Byrd (1998) presentó la integración de varios recursos estructurados de conocimiento léxico derivado a partir de diccionarios electrónicos monolingües y bilingües

[1] International Press Telecomunication Council (IPTC) se dedica a establecer estándares y formatos para la transmisión de información entre las agencias de noticias. Para obtener más información del sistema de clasificación IPTC acceder a la dirección `http://www.iptc.org`.

y tesauros. Rigau (1994) presentó un método automático para enriquecer semánticamente el diccionario electrónico español, mediante el uso de unas fórmulas de distancia conceptual muy simples junto con un diccionario bilingüe español/inglés y WordNet. Yarowsky (1992) propuso un método estadístico para enriquecer las definiciones de los diccionarios con las categorías del tesauro Roget. Este método fue aplicado por Rigau *et al.* (1998) para etiquetar un diccionario electrónico español monolingüe con categorías de WordNet. También, Chen y Chang (1998) usaron el tesauro Roget para etiquetar LDOCE. Magnini y Cavaglia (2000) presentaron una versión aumentada de la parte nominal de WordNet, anotando los synsets semi-automáticamente con una o más categorías.

Los recursos léxicos son componentes fundamentales en los sistemas de Procesamiento del Lenguaje Natural, tales como Recuperación de Información, Extracción de Información, Traducción Automática, Clasificación de Documentos, etc. Estos últimos requieren grandes lexicones en los que se organicen las palabras más relevantes a cada dominio semántico o categoría donde queramos clasificar un documento. Por ejemplo, la categoría *HEALTH* puede estar formada por las palabras *doctor*, *hospital*, *to operate*, etc. O la categoría *ECONOMY* puede estar formada por *money*, *stock*, *to invest*, *to buy*, etc. Lexicones de este tipo pueden obtenerse de Tesoros de la lengua (tesauros) o diccionarios ideológicos, donde se clasifican las palabras por categorías o dominios. Entre ellos, seguramente el más utilizado en PLN ha sido el *Roget´s Thesaurus* (Chapman, 1984).

Por su parte, los sistemas de clasificación intentan estructurar y organizar el conocimiento y la información. Así, en Biblioteconomía, los sistemas de clasificación (por ejemplo, el sistema *Library of Congress Classification*[2] o *Dewey Decimal Classification*[3]) permiten organizar todas las referencias bibliográficas. O en las agencias de noticias y periódicos permiten organizar sus noticias y artículos en secciones (por ejemplo, el sistema IPTC).

[2] `http://lcweb.loc.gov/catalog`

[3] `http://www.noblenet.org/wakefield/rdewey.htm`

O en los buscadores de Internet permiten dar estructura temática a la web.

Por otra parte, WordNet, seguramente el lexicon del inglés más utilizado en PLN, presenta una división de los sentidos de las palabras con demasiado detalle. Por ejemplo, a la palabra en inglés *line* le asocia 29 sentidos distintos. Por eso, las categorías o etiquetas de dominio, como *Agriculture, Health, etc*, aportan una forma más natural para establecer distinciones claras entre los sentidos de las palabras.

En el presente capítulo se define y describe un método automático para enriquecer semánticamente WordNet versión 1.6. con las categorías utilizadas en el sistema de clasificación de noticias IPTC[4], aplicando el método Marcas de Especificidad propuesto en esta Tesis.

5.1.1 IPTC Subject Reference System

El sistema de clasificación IPTC Subject Reference System ha sido desarrollado para que las agencias de noticias puedan codificar los artículos de prensa utilizando un idioma común.

IPTC Subject Reference System está compuesto de 17 categorías principales, y estas se subdividen a su vez en otras categorías hasta un tercer nivel. Estas categorías están subdivididas en grupos de palabras fuertemente relacionadas a ella. Las categorías principales de IPTC y las subcategorías que contienen cada una se muestran en la Tabla 5.1.

Para la aplicación del método propuesto se han considerado también como principales las subcategorías pertenecientes a la categoría principal *(Economy, Business y Finance)*. Estas categorías comentadas se muestran en la Tabla 5.2.

5.2 Método para enriquecer WordNet

A continuación se describen brevemente los procesos utilizados para enriquecer la base de datos léxica WordNet versión 1.6. La

[4] Este trabajo ha sido realizado dentro del marco del proyecto NAMIC IST-1999-12302 (`http://namic.itaca.it`).

Categorías Principales	N° de Palabras
Arts, Culture y Entertainment	23
Crime, Law y Justice	8
Disasters y Accidents	10
Economy, Business y Finance	130
Education	8
Environmental Issues	11
Health	12
Human Interest	6
Labour	18
Lifestyle y Leisure	18
Politics	27
Religion y Belief	7
Science y Technology	10
Social Issues	20
Sport	81
Unrest, Conflicts y War	12
Weather	6

Tabla 5.1. Categorías principales de IPTC

Subcategorías	N° Palabras
Agriculture	6
Chemicals	9
Computing y Technology	10
Construction y Property	5
Energy y Resources	14
Financial y Business Services	13
Consumer Goods	10
Macro Economics	12
Markets y Exchanges	6
Media	12
Metal Goods y Engineering	9
Metals y Minerals	5
Tourism y Leisure	7
Transport	4

Tabla 5.2. Categorías del nivel 2 de *(Economy, Business y Finance)*

figura 5.1 ilustra el proceso seguido para enriquecer WordNet. En primer lugar, el grupo de palabras (nombres para IPTC) pertenecientes a cada una de las categorías del sistema de clasificación IPTC forman el fichero de entrada al módulo WSD. Este módulo de WSD, basado en el método de Marcas de Especificidad, consulta la base de conocimiento WordNet para cada una de las palabras que forman parte de la categoría semántica de entrada, y devuelve para cada una de ellas el sentido correcto asignado por WordNet en un nuevo fichero. Este nuevo fichero obtenido es la entrada al modulo de reglas. Este módulo aplicará un conjunto de reglas para encontrar el super-concepto[5] en WordNet. Una vez encontrado este, se etiqueta en WordNet con su categoría correspondiente del sistema de clasificación IPTC, así como todos sus hipónimos.

En las siguientes sub-secciones se detallarán cada uno de los procesos nombrados anteriormente para enriquecer y etiquetar WordNet con categorías del sistema de clasificación IPTC.

5.2.1 Proceso 1: Obtención y Tratamiento de categorías

El grupo de palabras (nombres para IPTC) pertenecientes a cada una de las categorías del sistema de clasificación IPTC forman la entrada de información a este proceso. La categoría completa *Health* del sistema de clasificación IPTC se muestra en la figura 5.2.

A veces, esta información de entrada conlleva algunos problemas debidos a que las categorías IPTC tienen palabras compuestas de dos o más palabras y es imposible encontrarla en WordNet, impidiendo al sistema poder asignarle un sentido correcto. Por ejemplo, la categoría *Health* de IPTC tiene una palabra compuesta de dos palabras: *Health Organization*. Esta si se busca en WordNet no se encuentra y por lo tanto no se le asignaría la etiqueta del sentido WordNet correspondiente. Con objeto de resolver este problema se aplica la utilidad de WordNet denominada *"Find Keywords by Substring (grep)"*. Esta utilidad tiene la característica de relacionar otros *synsets* de WordNet con las palabras

[5] Super-concepto es el *synset* de WordNet que agrupa los sentidos desambiguados de una categoría determinada.

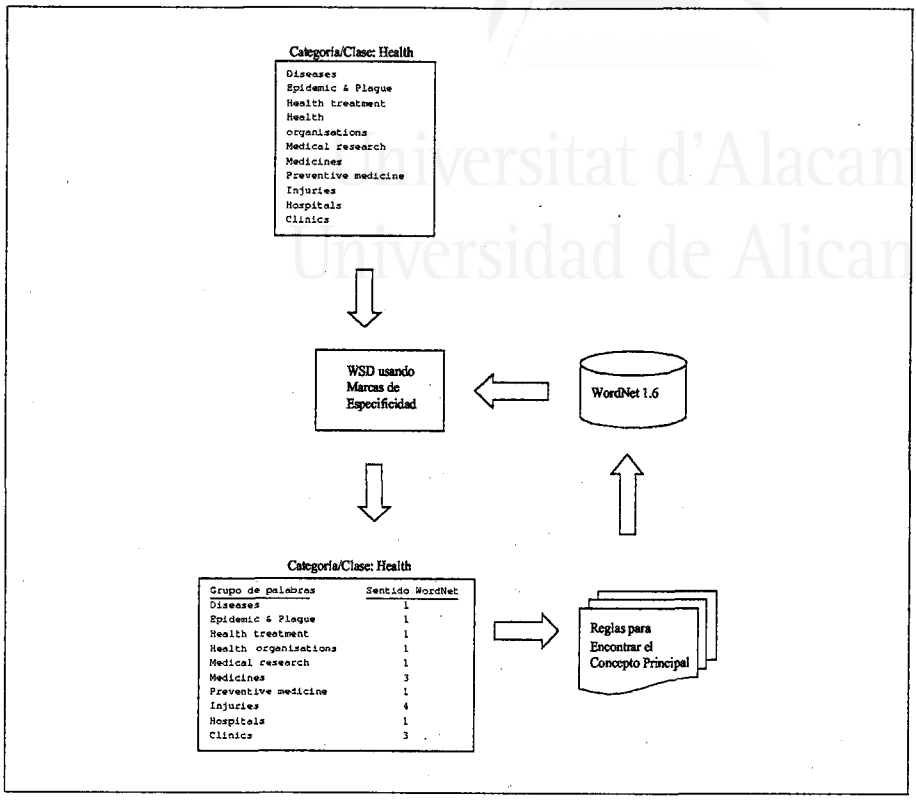

Figura 5.1. Proceso para enriquecer WordNet con categorías IPTC

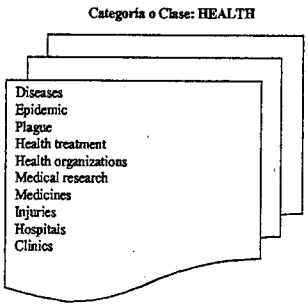

Figura 5.2. Categoría *Health* del sistema de clasificación IPTC

compuestas deseadas. Para el ejemplo anterior esta utilidad devolvería los *synset* de WordNet siguientes: *World health organization* y *Health maintenance organization.*

5.2.2 Proceso 2: Selección del sentido correcto

El núcleo principal de este proceso es el método de Marcas de Especificidad con heurísticas, explicado detalladamente en el capítulo 3. Aunque en la figura 5.3 se presenta la idea fundamental e intuitiva de este método.

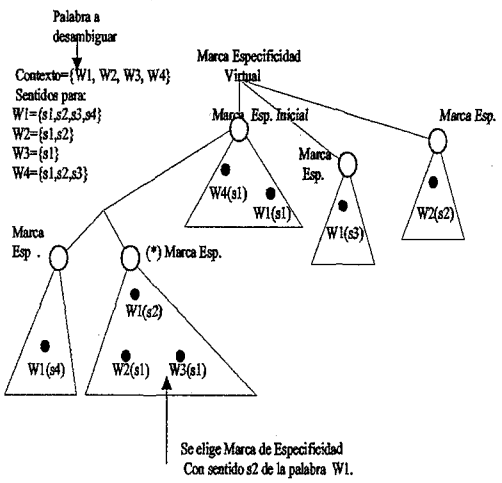

Figura 5.3. Idea intuitiva del método de Marcas de Especificidad

Este proceso toma como entrada al grupo de palabras pertenecientes a una de las categorías del sistema de clasificación IPTC tratadas en el proceso anterior. Y finalmente devuelve las mismas palabras de entrada etiquetadas con el número de sentido de WordNet mediante la aplicación del método de Marcas de Especificidad. Un ejemplo de la información devuelta por este proceso se muestra en la figura 5.4

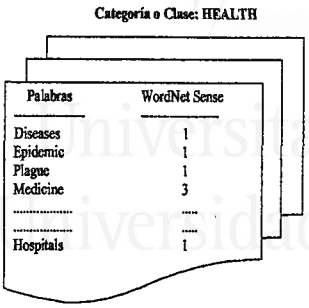

Figura 5.4. Sentidos asignados a la categoría *Health*

5.2.3 Proceso 3: Etiquetado del super-concepto

Este proceso tiene la función de analizar y seleccionar los super-conceptos de las palabras pertenecientes a la categoría tratada, usando las relaciones jerárquicas en la taxonomía de WordNet. Por ejemplo, en la figura 5.5 se muestra un ejemplo donde el super-concepto para la palabra *disease#1* es *ill_health#1*, porque *disease#1* es un hipónimos de *ill_health#1*.

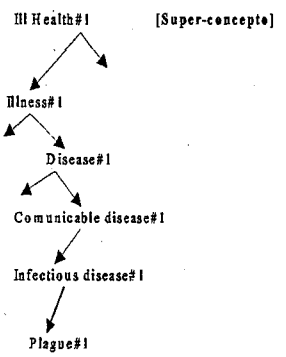

Figura 5.5. Ejemplo para la obtención del super-concepto

Este proceso toma como entrada al grupo de palabras con sus posibles sentidos etiquetados, pertenecientes a una de las cate-

gorías del sistema de clasificación IPTC. Al estar estas palabras etiquetadas con el número de sentido correspondiente a WordNet se tiene también el número de *synset* y su posición en la jerarquía, conociéndose el resto de *synsets* hiperónimos e hipónimos. Por lo tanto a partir de esta información de entrada se analiza el modo de combinar las categorías semánticas del sistema de clasificación IPTC y WordNet con el objetivo de obtener el super-concepto de WordNet para cada una de las palabras que pertenecen a la categoría semántica tratada. Para realizar esta tarea se aplican cuatro reglas una detrás de otra, las cuales se definen en las sub-secciones siguientes.

Regla 1. Si un *synset* determinado contiene solo como palabras hipónimas algunas pertenecientes a la categoría tratada, este se elige como super-concepto. Es decir, esta regla selecciona el super-concepto más bajo común a tantos *synsets* desambiguados. Por ejemplo, la categoría *health* está compuesta de un grupo de palabras y entre estas se incluyen la palabras *clinic* y *hospital*. Estas a su vez han sido desambiguadas con el sentido *clinic#3* y *hospital#1*. Por lo tanto, al analizar estos dos *synsets* se observa que son hipónimos de *medical_building#1* y como este tiene solamente como hipónimos a los dos anteriores entonces se elige como super-concepto. La figura 5.6 ilustra el ejemplo anteriormente comentado.

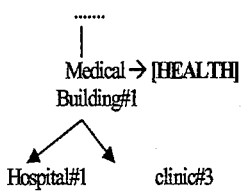

Figura 5.6. Ejemplo de aplicación de la Regla 1

Regla 2. Si el *synset* desambiguado tiene un hiperónimo que está compuesto por una palabra idéntica a la categoría, entonces este *synset* hiperónimo se elige como super-concepto. Por ejemplo, en-

tre las palabras de la categoría *health* se incluye la palabra *disease*. Esta a su vez ha sido desambiguada con el sentido *disease#1*. Al analizar este *synset*, recorriendo la jerarquía de hiperónimos, se observa que el *synset ill_health* es hiperónimo de *disease#1* dos niveles hacia arriba. Por lo tanto, al aplicar esta regla se observa que el *synset ill_health* está compuesto de la palabra *ill* y de *health*, siendo esta última la misma palabra de la categoría tratada, y como consecuencia se elige a *ill_health* como super-concepto. La figura 5.7 ilustra el ejemplo anteriormente comentado.

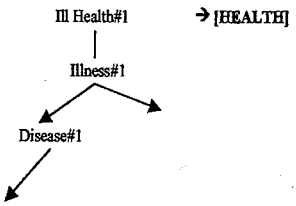

Figura 5.7. Ejemplo de aplicación de la Regla 2

Regla 3. Esta regla resuelve el problema de aquellas palabras que ni están relacionadas directamente en WordNet ni están en algún synset compuesto de las relaciones de hiperonimia e hiponimia. Para resolver este problema se ha usado la glosa de cada *synset* de las relaciones de hiponimia de WordNet.

En primer lugar, para cada palabra perteneciente a la categoría a tratar de IPTC se eligen sus distintos sentidos o *synsets* en la jerarquía de WordNet. Para estos se elige el *synset* hiperónimo inmediato y posteriormente son analizadas todas las glosas pertenecientes a los *synsets* hipónimos y la suya propia, con el objetivo de comprobar si alguna palabra perteneciente a la categoría semántica es encontrada en dichas glosas. En caso de que se encuentren palabras de la categoría en las glosas de esos *synsets* hipónimos se seleccionan para su posterior etiquetación.

Por ejemplo, entre las palabras de la categoría *HEALTH* se incluye la palabra *hospital*. Según el contexto asociado a esta

palabra y aplicando el proceso 2, se etiqueta *hospital#1* como sentido correcto. Pero, al aplicar esta regla, el otro sentido *hospital#2* tiene como *synset* hiperónimo a *medical_institution#1*, el cual es definido por su glosa como *"an institution created for the practice of medicine"*. Como se observa, en su definición aparece la palabra *"medicine"* y esta es una de las palabras que también pertenece a la categoría estudiada, por lo tanto estará ralacionada semánticamente. Si se continua con todas las glosas de sus *synsets* hipónimos se encuentra a *hospital#2* y a *clinic#1*, y estas a su vez también tienen palabras en sus glosas que pertenecen a la categoría estudiada como *medical*. Como consecuencia de esto se decide que el *synset medical_institution#1* y todos sus hipónimos se etiqueten con la categoría semántica *HEALTH*. La figura 5.8 ilustra el ejemplo anteriormente comentado.

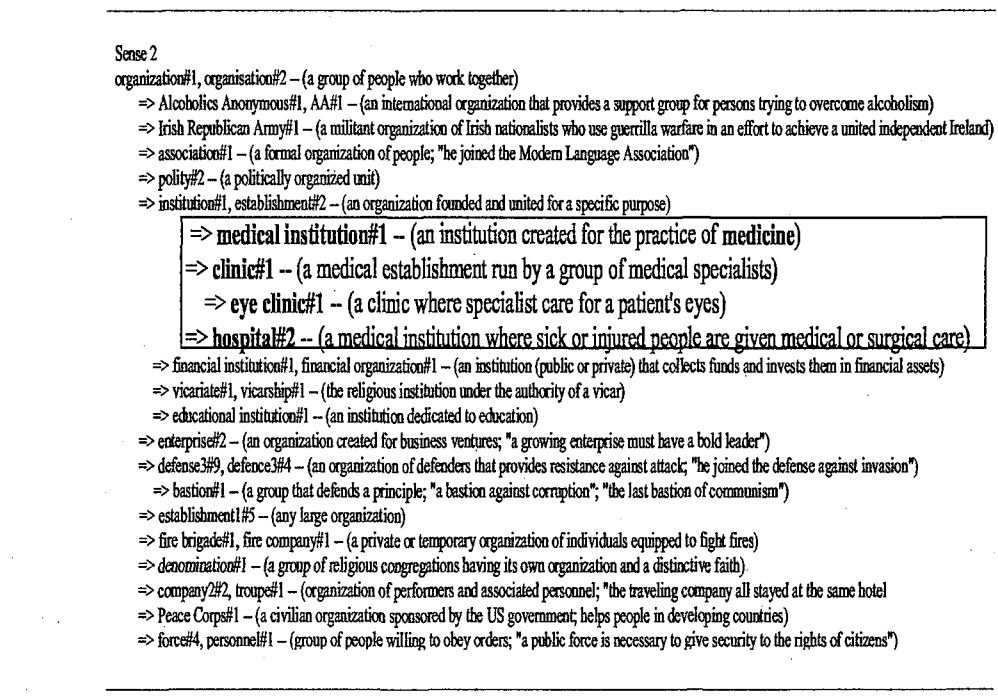

Figura 5.8. Ejemplo de aplicación de la Regla 3

Regla 4. Si un *synset* desambiguado está próximo al nivel raíz de la jerarquía WordNet (segundo o tercer nivel), es decir próximo a la cima de la taxonomía, se le etiqueta con la categoría semántica sin más. Esto es debido a que estos *synsets* son muy genéricos y sus distintos hipónimos estarán poco relacionados entre sí. Por ejemplo, la categoría *HEALTH* está compuesta de un grupo de palabras y entre estas se incluye la palabra *epidemic*. Según el contexto asociado a esta palabra y aplicando el proceso 2, se etiqueta *epidemic#1* como sentido correcto. Este está en el nivel 4 de la jerarquía semántica de WordNet, por lo que su *synset* hiperónimo es *outbreak#1*. Su significado, " una repentina y violenta aparición de una condición indeseable", representa algo mucho más genérico que lo que realmente significa una "epidemia". Por lo tanto, se etiqueta con la categoría semántica *HEALTH* solamente a *epidemic#1*. La figura 5.9 ilustra el ejemplo anteriormente comentado.

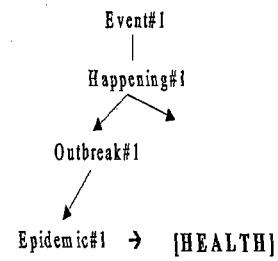

Figura 5.9. Ejemplo de aplicación de la Regla 4

5.2.4 Proceso 4: Etiquetado por expansión

Una vez hallados los super-conceptos de una categoría IPTC, el proceso 4 se encarga de asignar la etiqueta de la categoría a los *synsets* hipónimos y merónimos de los super-conceptos.

Este proceso toma como entrada el conjunto de super-conceptos y la categoría tratada. Posteriormente recorre la jerarquía de WordNet a través de las relaciones de hiponimia y meronimia

y asigna a cada uno de estos *synsets* la etiqueta de la categoría. Al concluir este proceso, la base de datos de WordNet se enriquece con las etiquetas de las categorías IPTC en cada *synset* hipónimo y merónimo de los super-conceptos obtenidos en los procesos anteriores. Un ejemplo de los super-conceptos con su etiqueta se muestra en la figura 5.10.

WordNet Sense Word	Super-Concept
{10129713} disease#1	{10120678} <IPTC.Health> ill Health#1
{05531449} epidemic#1	{05531449} <IPTC.Health> epidemic#1
{10169961} plague#1	{10120678} <IPTC.Health> ill Health#1
{00430183} treatment#1	{00430183} <IPTC.Health> treatment#1
{00650412} organisation#6	{00650412} <IPTC.Health> organisation#6
{10366424} health#1	{10366424} <IPTC.Health> health#1
{00092823} medical#1	{00092823} <IPTC.Health> medical#1
{02981307} medicine#2	{02981307} <IPTC.Health> medicine#2
{10257548} injury#1	{10120678} <IPTC.Health> ill Health#1
{02834408} hospital#1	{02980771} <IPTC.Health> medical_building#1
{02450085}clinic#3	{02980771} <IPTC.Health> medical_building#1
{10129713} preventative#1	{02981307} <IPTC.Health> medicine#2

Figura 5.10. Etiquetado de la categoría *Health*

En esta figura solamente se muestran los super-conceptos etiquetados con la categoría, sin embargo también estarían etiquetados todos los *synsets* hipónimos y merónimos de cada uno de estos super-conceptos. Con esto se enriquecería y ampliaría la base de conocimiento WordNet con etiquetas del sistema de clasificación IPTC.

5.3 Evaluación del proceso de enriquecimiento de WordNet

En esta sección se describen los experimentos realizados y los resultados obtenidos al aplicar el método descrito en las secciones anteriores, usando la interfaz desarrollada para este propósito. Con estos experimentos lo que se pretende es evaluar la efectividad

del método propuesto para enriquecer WordNet con categorías del sistema de clasificación IPTC.

La experimentación del presente método se dividió en dos experimentos.

5.3.1 Experimento 1: Evaluación del método de Marcas de Especificidad con Sistemas de Clasificación

El objetivo de este experimento es evaluar si el método de Marcas de Especificidad desambigua efectivamente el sentido de las palabras que forman parte de las categorías en los sistemas de clasificación. En definitiva, lo que se pretende con este experimento es confirmar que el sistema de Marcas de Especificidad obtiene muy buenos resultados con palabras fuertemente relacionadas, como es el caso de los sistemas de clasificación.

Los resultados obtenidos en la evaluación del método de Marcas de Especificidad sobre los grupos de palabras relacionadas por las categorías descritas en la sección anterior se muestran en la Tabla 5.3. En ella se presentan las medidas de "precisión", "cobertura" y "cobertura absoluta", así como los resultados obtenidos en el experimento 2 del capítulo 4 al aplicar el método sobre textos no restringidos (*SemCor y Encarta 98*), con el objetivo de mostrar la gran diferencia entre unos valores y otros.

Tipo de Texto	Nombre del Corpus	Ratio	Valores
Textos No Restringidos	SEMCOR	Precisión	67,2 %
		Cobertura	66,2 %
		C. absoluta	98,5 %
	ENCARTA	Precisión	67 %
		Cobertura	66,3 %
		C. absoluta	98,8 %
Sistema de Clasificación	IPTC	Precisión	96.1 %
		Cobertura	92,5 %
		C. absoluta	96,8 %

Tabla 5.3. Resultados de "precisión", "cobertura" y "Cobertura absoluta" sobre IPTC

Como se puede observar en la Tabla 5.3, se obtienen unos resultados del 96,1 % de "precisión" y del 92,5 % de "cobertura", cuando se aplica el método a todas las categorías IPTC (30 categorías) y a 399 palabras pertenecientes a dichas categorías. Estos resultados son muy superiores a los obtenidos en el experimento 2 del capítulo 4 y mostrados en la misma tabla, cuando se aplica el método de Marcas de Especificidad sobre textos no restringidos (67,2 % de "precisión" para *SemCor* y 67 % de "precisión" para *Encarta 98*). Esta diferencia se debe a que el método de Marcas de Especificidad se basa en el conocimiento de cuantas palabras del contexto se agrupan alrededor de una clase semántica. Por lo tanto, las conclusiones que se obtienen de este experimento son que el método de Marcas de Especificidad aplicado funciona con una efectividad muy alta sobre palabras fuertemente relacionadas y agrupadas en una única categoría, debido a que esas palabras pertenecientes a las categorías están relacionadas muy fuertemente y WordNet también refleja esa relación fuerte en su taxonomía.

5.3.2 Experimento 2: Evaluación del método de Enriquecimiento de WordNet

El segundo experimento tiene como objetivo evaluar el enriquecimiento de WordNet con las categorías IPTC.

El primer paso del experimento se basa en calcular la cantidad de *synsets* de WordNet correctamente etiquetados, incorrectamente etiquetados y no etiquetados (no están en WordNet) para el conjunto de palabras pertenecientes a las categorías IPTC. En la Tabla 5.4 se muestran los resultados obtenidos al aplicar el proceso anterior. En dicha tabla la columna de palabras indica las palabras junto con las sub-categorías.

A continuación, con el objetivo de evaluar la "precisión", "cobertura" y "cobertura absoluta" del método, se aplican las cuatro reglas descritas anteriormente y se comprueba manualmente los resultados obtenidos para cada una de las palabras pertenecientes a las categorías IPTC. Para este experimento la "precisión" se calcula mediante el ratio entre los *synsets* correctamente etiquetados y los *synsets* contestados (correctamente e incorrectamente

Categorias IPTC	Palabras	Correc.	Incorrec	Sin etiq.
Arts, culture & entertainment	23	21	2	0
Crime, law & Justice	8	8	0	0
Disasters & accidents	10	7	2	1
Agriculture	6	5	0	1
Chemical	9	8	0	1
Computing & Technology	10	9	1	0
Construction & property	5	3	1	1
Energy & resource	14	10	0	4
Financial & business services	13	12	1	0
Consumer goods	10	10	0	0
Macro economics	12	10	0	2
Markets & exchanges	6	6	0	0
Media	12	12	0	0
Metal goods & engineering	9	8	0	1
Metals & minerals	5	5	0	0
Tourism & leisure	7	7	0	0
Transport	4	4	0	0
Education	8	6	0	2
Environmental Issues	11	11	0	0
Health	12	8	3	1
Human Interest	6	6	0	0
Labour	18	17	1	0
Lifestyle & leisure	18	18	0	0
Politics	27	22	3	2
Religion & belief	7	7	0	0
Science & technology	10	9	0	1
Social Issues	20	19	1	0
Sport	81	72	1	8
Unrest, conflicts & war	12	12	0	0
Weather	6	6	0	0
TOTAL	399	358	16	25

Tabla 5.4. Categorías IPTC con los resultados obtenidos

etiquetados). "cobertura" se calcula mediante el ratio entre los *synsets* correctamente etiquetados y el número total de palabras. Y "cobertura absoluta" se calcula mediante el ratio entre los *synsets* contestados (correctamente e incorrectamente etiquetados) y el número total de palabras. Los resultados de "precisión", "co-

bertura" y "cobertura absoluta" del método se muestran en la Tabla 5.5.

%	C. absoluta	Precisión	Cobertura
Enriquecimiento WordNet	93.7 %	95.7 %	89.8 %

Tabla 5.5. Resultados obtenidos de "precisión", "cobertura" y "cobertura absoluta"

Una observación a los resultados obtenidos es que si el método de Marcas de Especificidad desambigua correctamente y se aplican las reglas presentadas anteriormente, el método propuesto funciona correctamente. Sin embargo, si el método de Marcas de Especificidad desambigua incorrectamente, el método etiqueta incorrectamente a los *synsets* de WordNet con las categorías de IPTC. Pero, la conclusión obtenida al observar los resultados presentados en la Tabla 5.5 del experimento es que el método aquí utilizado para enriquecer WordNet es preciso, eficiente y eficaz cuando se aplica a IPTC *Subject Reference System*.

5.4 Implementación de la interfaz para enriquecer WordNet

En esta sección se presenta la interfaz implementada para enriquecer WordNet con categorías semánticas de sistemas de clasificación. Esta interfaz está compuesta de un conjunto de programas diseñados para realizar las especificaciones descritas en la sección anterior 5.2. Estos programas realizan automáticamente cada uno de los procesos explicados así como el etiquetado de los *synsets* en WordNet.

A continuación se describen las características del diseño e implementación de la interfaz con el objetivo de extender y mejorar la base de conocimiento léxica WordNet.

El diseño de la interfaz está compuesta de los cuatro procesos explicados detalladamente en la sección 5.2. Pero para una mejor comprensión se presenta la figura 5.11 en donde se resume de forma esquemática el diseño utilizado para la implementación.

Hay que aclarar que esta interfaz está preparada para trabajar con cualquier sistema de clasificación, pero en el presente estudio solo se ha trabajado sobre IPTC.

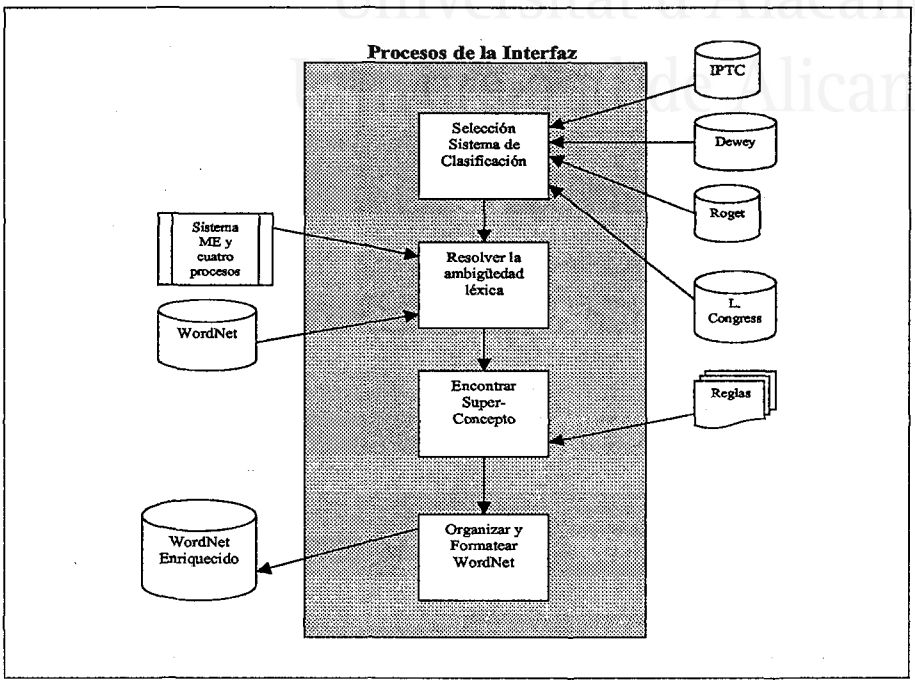

Figura 5.11. Proceso que intervienen en la interfaz

Esta interfaz está programada con el lenguaje de programación C++ y ofrece al usuario las siguientes operaciones.

- **Elegir la categoría a procesar.** Una vez seleccionada la categoría, el grupo de palabras pertenecientes a ella aparecen en la ventana izquierda de la interfaz.
- **Etiquetar.** Cuando el usuario elige este comando se activan los procesos de resolución de la ambigüedad léxica así como la elección del super-concepto. La salida de información, correspondiente al grupo de palabras de la categoría seleccionada suministrada por estos dos procesos, aparece en la ventana derecha de la interfaz denominada "*synsets* etiquetados". Esta salida de información está compuesta de las palabras con el sen-

tido de WordNet y el super-concepto asignado a cada palabra perteneciente a la categoría previamente elegida. En la figura 5.12 se muestra la información que se visualiza en la ventana denominada "*synsets* etiquetados".

- **Almacenar y etiquetar categorías en WordNet.** Cuando esta opción es elegida, toda la información obtenida en los procesos previos es organizada, formateada y almacenada para cada super-concepto y todos sus *synsets* hipónimos y merónimos en la base de datos léxica WordNet.

Con objeto de aclarar cada uno de los procesos implementados y mostrados en el diseño de la figura 5.11, a continuación, se presenta mediante la figura 5.12 la interfaz implementada para enriquecer WordNet con categorías semánticas de sistemas de clasificación

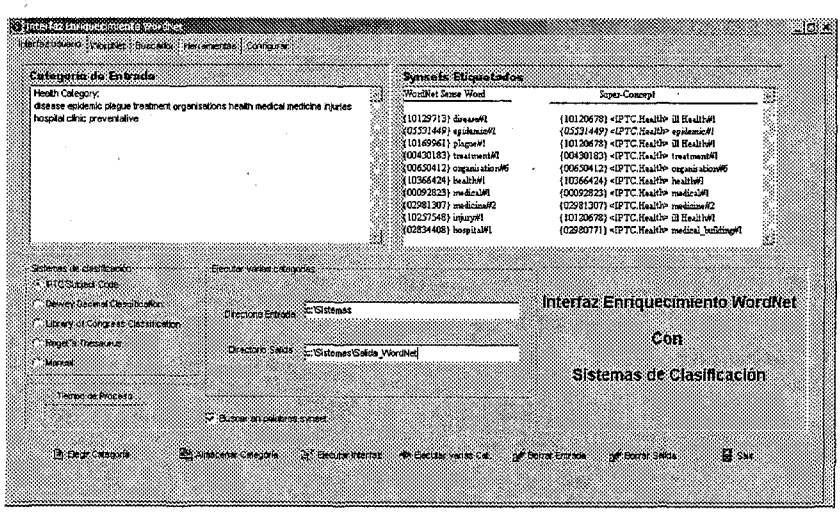

Figura 5.12. Interfaz de Usuario para enriquecer WordNet con categorías de sistemas de clasificación

5.5 Conclusiones y aportaciones obtenidas

Las conclusiones obtenidas en este capítulo son que el método de Marcas de Especificidad se aplica satisfactoriamente para asignar categorías del sistema de clasificación IPTC *Subject Reference System* a la base de datos léxica WordNet, con objeto de enriquecer su taxonomía. También, los resultados de los experimentos realizados cuando el método se aplica a IPTC *Subject Reference System* indican que este método para enriquecer WordNet es preciso y eficaz.

De los experimentos realizados debemos destacar:

- En primer lugar, el método de Marcas de Especificidad funciona satisfactoriamente con sistemas de clasificación, ya que sus categorías están subdivididas por grupos de palabras que están fuertemente relacionadas. Y esto facilita la desambiguación del sentido de estas palabras debido a que el método de Marcas de Especificación se basa en las relaciones de hiperonimia e hiponimia en la taxonomía de WordNet. Aunque el sistema ha sido probado sobre el sistema de clasificación IPTC *Subject Reference System*, este se puede aplicar a otros sistemas que agrupen palabras sobre una categoría como por ejemplo *Library of Congress Classification Roget´s Thesaurus* o *Dewey Decimal Classification.*
- En segundo lugar, la reducción de las palabras polisémicas es otra consecuencia relevante de aplicar el método para enriquecer WordNet. Porque el número de categorías para una palabra es normalmente mas bajo que el número de sentidos para esa palabra. Por ejemplo, las categorías *Health, Sports, etc* establecen una forma de agrupar los sentidos de las palabras. Por lo tanto, este método puede ayudar a reducir el problema de la "granularidad fina" en la distinción de los sentidos en WordNet (Ide y Véronis, 1998).
- Finalmente, los investigadores, gracias a esta aportación, serán capaces de construir variantes de sistemas WSD, ya que estos ahora para desambiguar una palabra de un texto elegirán la etiqueta de su categoría en vez de elegir la etiqueta de su sentido (Magnini y Strapparava, 2000).

6. Conclusiones y trabajos futuros

En este trabajo se ha desarrollado un método para resolver la ambigüedad léxica pura (semántica) en textos de dominios no restringidos en cualquier lengua que tenga un WordNet[1] particular (inglés, español, italiano, etc). El método de resolución de la ambigüedad léxica pura propuesto se basa en el uso de conocimiento lingüístico (información léxica y morfológica) y de conocimiento a partir de las relaciones léxicas y semánticas de un recurso externo (taxonomía de nombres a partir de la base de datos léxica WordNet), ambos independientes del dominio.

Gracias al uso de información no dependiente del dominio, este método de resolución de la ambigüedad léxica, al que hemos denominado Marcas de Especificidad, está preparado para ser aplicado a sistemas que trabajen sobre cualquier dominio y sobre cualquier lengua que tenga un WordNet predefinido.

A continuación se presentan las principales aportaciones de esta Tesis, así como los posibles trabajos futuros y la producción científica obtenida en la realización de este trabajo.

6.1 Aportaciones

Las principales aportaciones de esta Tesis han sido:

- La definición de la noción de Marca de Especificidad, la cual formaliza la relación entre sentidos que se basan en taxonomías. La Marca de Especificidad se aprovecha de la información suministrada por las relaciones de sinonimia, hiperonimia e hiponimia. Aunque en esta Tesis se ha utilizado para nombres, se

[1] Si cada WordNet particular tiene definiciones es mejor para el funcionamiento del método.

puede adecuar para trabajar con verbos. Con la utilización de las Marcas de Especificidad se aporta una noción como base para la desambiguación del sentido de las palabras, los cuales están unidos a través de conceptos en las taxonomías. Además, la aplicación de esta noción no requiere desambiguación manual previa para aplicarse.

- La definición de un método basado en las Marcas de Especificidad, el cual utiliza el paradigma de los métodos basados en conocimiento, y en concreto el conocimiento que suministra la base de conocimiento léxica WordNet. Debido a las características de las Marcas de Especificidad, se ha diseñado un método que desambigua nombres de un texto libre según estén organizados los sentidos de las palabras en la taxonomía. Por tal motivo, este método puede aplicarse a textos de diferentes idiomas sin realizar adaptaciones del mismo, siempre que tenga un WordNet.
- La definición de un conjunto de heurísticas para mejorar la propuesto del método inicial, que denominamos método Marcas de Especificidad. Estas heurísticas utilizan el conocimiento suministrado por WordNet, como son el *synset* y la glosa en su idioma, por lo tanto también pueden aplicarse a textos de diferentes idiomas sin realizar adaptaciones.
- El diseño y desarrollo de una interfaz para implementar el método Marcas de Especificidad. Esta interfaz puede ser ejecutada desde cualquier ordenador que tenga acceso a Internet y que acceda a la dirección `http://gplsi.dlsi.ua.es/wsd`.
- El entrenamiento y ajuste del método Marcas de Especificidad implementado. Para ello el método se ha validado y ajustado con el objetivo de estudiar su efectividad mediante un trabajo experimental dividido en tres experimentos. El experimento 1 del capítulo *Experimentación* consiste en comprobar el funcionamiento del método cuando se aplica solamente la noción de Marca de Especificidad. El experimento 2 consiste en comprobar que al complementar el método con un conjunto de heurísticas, estas aportan mejores porcentajes de desambiguación y por lo tanto mejoran el método. Y el experimento 3 consiste en analizar y definir la ventana óptima de contexto para obtener la

mejor desambiguación. Con estos tres experimentos se consigue que el método desambigue los sentidos de las palabras sobre WordNet con una "precisión" del 67,1 %, una "cobertura " del 66,2 % y una "cobertura absoluta" del 98,5 %.

- La comparación del método de Marcas de Especificidad con otros métodos de WSD y su evaluación final. Se ha realizado una comparación *directa* e *indirecta* del método mediante dos experimentos. En el experimento 4 del capítulo *Experimentación* se realiza una comparación con métodos basados en el conocimiento, es decir métodos WSD que pertenecen a la misma clasificación que el propuesto en esta Tesis. Y en el experimento 5 se realiza una comparación entre el método propuesto y uno basado en corpus, concretamente en un modelo probabilístico que utiliza el principio de Máxima Entropía. A partir de esta comparación y comprobar los resultados obtenidos, podemos probar que el método de Marcas de Especificidad es útil para desambiguar el sentido de las palabras, ya que se han obtenido mejores resultados que otros métodos basados en conocimiento, y en concreto el conocimiento que suministra la base de conocimiento léxica WordNet.

  Para la evaluación final del método Marcas de Especificidad se han realizado dos experimentos. El experimento 6 consiste en participar con el método en la competición de sistemas de WSD denominada SENSEVAL-2 para los nombres seleccionados por el comité en la tarea de "lexical sample" tanto para inglés como para español. Y el experimento 7 consiste en evaluar al método cuando se aplica con las heurísticas activadas en cascada o secuencialmente y cuando se aplican independientemente unas de otras, y en ambos casos sobre todos los documentos del corpus *SemCor*.
- La definición de un método para enriquecer semánticamente el recurso léxico WordNet con categorías o clases de otros sistemas de clasificación, mediante la aplicación del método de Marcas de Especificidad. El sistema de clasificación utilizado para etiquetar automáticamente los *synsets* de WordNet ha sido IPTC Subject Reference System.

El recurso léxico WordNet, también muy utilizado en PLN, presenta una división de los sentidos de las palabras con demasiado detalle (en inglés conocido como fine-grained). Por eso, las categorías, como *Agriculture, Health, etc*, aportan una forma más adecuada para diferenciar los sentidos de las palabras. Por lo tanto para tratar y resolver este problema asociado a WordNet, se define y describe un método automático para enriquecer semánticamente WordNet versión 1.6. con las categorías utilizadas en el sistema de clasificación IPTC.

- El diseño y desarrollo de la interfaz construida para extender y mejorar la base de datos léxica WordNet con categorías del sistema de clasificación. Gracias al desarrollo de esta interfaz se han podido realizar los experimentos para evaluar la efectividad del método propuesto para enriquecer WordNet. Como resultado se obtuvo una "cobertura absoluta" del 93.7 %, una "precisión" del 95.7 % y una "cobertura" del 89.8 %.

Como conclusión final indicaremos que en este trabajo hemos estudiado dos métodos, uno aplicado a la desambiguación del sentido de las palabras para textos no restringidos en cualquier idioma que tenga un WordNet particular y otro aplicado al enriquecimiento de WordNet con categorías de sistemas de clasificación. En ambos métodos los resultados obtenidos son satisfactorios y consideramos que el método de Marcas de Especificidad es útil para WSD y el de enriquecimiento de WordNet es útil para mejorar y extender la base de datos léxica WordNet con categorías de los sistema de clasificación.

6.2 Trabajos futuros

Como trabajo futuro se pretende modificar el algoritmo para utilizar más y mejor las relaciones semánticas de WordNet, además de añadir más categorías léxicas a la hora de desambiguar, como los verbos y los adjetivos. Esto hará que se tenga más información de contexto y mejor relacionado. A consecuencia de lo dicho anteriormente, también se pretende obtener más información de otros recursos léxicos como los diccionarios o tesauros. Pero quizá el

cambio más importante sea introducir más heurísticas que combinen diferentes técnicas y recursos para producir todos juntos una mejor desambiguación del sentido de las palabras.

Magnini y Strapparava (2000) estudiaron la influencia de los dominios en la desambiguación de los sentidos. Así, para cada palabra en un texto se elige una etiqueta de un dominio en vez de la etiqueta del sentido. Una consecuencia de aplicar los dominios obtenidos de WordNet es que se reduce la polisemia de las palabras, ya que el número de dominios para una palabra es generalmente más bajo que el número de sentidos para la misma. Por lo tanto, otro trabajo futuro sería modificar y adaptar nuestro método para obtener esta variante de WSD que se le conoce como *Word Domain Disambiguation* (WDD).

6.3 Producción científica

El trabajo de investigación llevado a cabo para la realización de esta Tesis ha tenido como consecuencia la publicación de diferentes publicaciones relacionadas con la resolución de la ambigüedad léxica pura y con el enriquecimiento de WordNet con categorías de los sistemas de clasificación. A continuación se presentan cada una de las publicaciones referentes a lo comentado anteriormente:

- Publicaciones en revistas nacionales
  - Suárez, A., Montoyo A. Estudio de cooperación de métodos de desambiguación léxica: marcas de especificidad vs. máxima entropía. *Procesamiento del Lenguaje Natural*, **27**. 2001. pp 207–214. ISSN:1135-5948.
  - Montoyo, A. Método basado en Marcas de Especificidad para WSD. *Procesamiento Lenguaje Natural*, **26**. 2000. pp 53–59. ISSN:1135-5948.
- Comunicaciones a congresos internacionales
  - Montoyo, A., Suárez, A. y Palomar M. Combining Supervised-Unsupervised Methods for Word Sense Disambiguation. *In Proceedings of the Third International conference on Intelligent Text Processing and Computational Linguistics (CICLing-*

2002). México D.F.(México). 2002. Lecture Notes in Computer Science, VOL 2276:156-164. Springer-Verlag.

- Montoyo A., Romero R., Vázquez S., Calle MC., Soler S. The Role of WSD for Multilingual Natural Language Applications. *Proceedings of 5th International Conference on Text Speech and Dialogue TSD'2002*. Brno (Czech Republic). 2002. Lecture Notes in Artificial Intelligence. Springer-Verlag (pendiente de publicar).
- Vázquez, S., Calle MC., Soler, S. y Montoyo, A. Specification Marks Method: Design and Implementation. *In Proceedings of the Third International conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*. México D.F.(México). 2002. Lecture Notes in Computer Science, VOL 2276:439-441. Springer-Verlag.
- Montoyo A., Suárez A. The University of Alicante Word Sense Disambiguation System. *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*. pp.131-134. Toulouse, France. July 2001.
- Montoyo A., Palomar M. y Rigau G. Method and Interface for WordNet Enrichment with Classification Systems. *Proceedings of 12th International Conference on Database and Expert Systems Applications DEXA-2001*. Munich, Germany. 2001. Lecture Notes in Computer Science, VOL 2113:122-130. Springer-Verlag.
- Montoyo A., Palomar M. y Rigau G. Method for WordNet Enrichment Using WSD. *Proceedings of 4th International Conference on Text Speech and Dialogue TSD'2001*. Selezná Ruda - Spièák, Czech Republic. 2001. Lecture Notes in Artificial Intelligence, VOL 2166:180-186. Springer-Verlag.
- Montoyo, A. y Palomar, M. Specification Marks for Word Sense Disambiguation: New Development. *Proceedings of the 2nd International conference on Intelligent Text Processing and Computational Linguistics (CICLing-2001)*. México D.F. (México). 2001. Lecture Notes in Computer Science, VOL 2004:182-191. Springer-Verlag.

- Montoyo A., Palomar, M. y Rigau, G. WordNet Enrichment with Classification Systems. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-01), Workshop WordNet and Other Lexical Resources: Applications, Extensions and Customisations.* Carnegie Mellon University. pp.101-106. Pittsburgh, PA, USA. 2001.
- Montoyo, A., Palomar M. y Rigau, G. Lexical enrichment of WordNet with Classification Systems using Specification Marks Method". *In Proceedings 6th International Conference on Application of Natural Language to Information Systems (NLDB´2001).* Madrid (Spain). 2001. Lecture Notes in Informatics, P-3:109-120, German Informatics Society GI-Edition.
- Montoyo, A. y Palomar M. WSD Algorithm applied to a NLP System. *In Proceedings of the 5th International Conference on Application of Natural Language to Information Systems (NLDB´2000).* Versailles (France). 2000. Lecture Notes in Computer Science, VOL 1959:54-65, Springer-Verlag.
- Montoyo, A. y Palomar M. Word Sense Disambiguation with Specification Marks in Unrestricted Texts. *In Proceedings of the 11th International Workshop on Database and Expert Systems Applications (DEXA 2000).* Greenwich, London. 2000. IEEE Computer Society. pp 103-108. ISBN 0-7695-0680-1.

• Informes técnicos

- LLopis, F., Muñoz, R., Suárez, A., Montoyo, A., Palomar, M., Ferrández, A., Mártinez-Barco, P., Peral, J., Romero, R., y Saiz, M. Propuesta de un sistema de extracción de información de textos notariales: EXIT. *Report Interno.* Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante. 1998.

Además de los anteriores, otros trabajos relacionados que han complementado los trabajos realizados para esta memoria son:

• Publicaciones en revistas nacionales

- Muñoz, R., Montoyo, A., Llopis, F., y Suárez, A. Reconocimiento de Entidades en el sistema EXIT. *Procesamiento Lenguaje Natural*, **23**:47-53, Septiembre 1998. ISSN:1135-5948.

- Llopis, F., Muñoz, R., Suárez, A., Montoyo, A. EXIT: Propuesta de un sistema de extracción de información de textos notariales". *Revista NOVATICA*, **133**:26-30. Mayo-Junio 1998.

- Comunicaciones a congresos internacionales
  - Muñoz R., Montoyo A. y Saiz-Noeda M.. Semantic Information in Anaphora Resolution. *In Proceedings of the International Conference Portugal for Natural Language Processing (PorTAL-2002)*. Universidade do Algarve, Faro, Portugal. 23 al 26 de junio. 2002. Lecture Notes in Computer Science, (pendiente de publicar). Springer-Verlag.
  - Soler, S. y Montoyo, A. A Proposal for WSD using Semantic Similarity. *In Proceedings of the Third International conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*. México D.F.(México). 2002. Lecture Notes in Computer Science, VOL 2276:165-167. Springer-Verlag.
  - Palomar M., Saiz-Noeda M., Muñoz, R., Suárez, A., Martínez-Barco, P., y Montoyo, A. PHORA: NLP System for Spanish. *In Proceedings 2nd International conference on Intelligent Text Processing and Computational Linguistics (CICLing-2001)*. México D.F.(México). 2001. Published in Lecture Notes in Computer Science, VOL 2004:126-139. Springer-Verlag.

Bibliografía

AGIRRE, E., X. ARREGI, X. ARTOLA, A. DÍAZ DE ILARRAZA y K. SARASOLA (1994). *Intelligent Dictionary Help Systems*, Applications and Implications of current LSP Research, Fakbokforlaget, Norway.

AGIRRE, E. y D. MARTÍNEZ (2000). «Exploring automatic word sense disambiguation with decision lists and the Web», en *Proceedings of the Semantic Annotation And Intelligent Annotation workshop organized by COLING*, Luxembourg.

AGIRRE, ENEKO y DAVID MARTINEZ (2001a). «Knowledge Sources for Word Sense Disambiguation», en *Proceedings of the 4th International Conference on Text, Speech and Dialogue (TSD-2001)*, págs. 1–10, Zelezna Ruda, Czech Republic.

AGIRRE, ENEKO y DAVID MARTINEZ (2001b). «Learning class-to-class selectional preferences», en *Proceedings of the Workshop "Computational Natural Language Learning" (CoNLL-2001). In conjunction with ACL'2001/EACL'2001*, Toulouse, France.

AGIRRE, ENEKO y GERMAN RIGAU (1995). «A proposal for Word Sense Disambiguation using Conceptual Distance», en *Proceedings of the International Conference "Recent Advances in Natural Language Processing" (RANLP´95)*, Tzigov Chark, Bulgaria.

AGIRRE, ENEKO y GERMAN RIGAU (1996). «Word Sense Disambiguation using Conceptual Density», en *Proceedings of the 16th International Conference on Computational Linguistic (COLING´96)*, Copenhagen, Denmark.

AHA, DAVID, DENNIS KIBLER y MARC ALBERT (1991). «Instance-based learning algorithms», *Machine Learning*, **6**(1),

37–66.

ALPAC (1966). «Language and Machine: Computers in Translation and Linguistics», *National Research Council Automatic Language Processing Advisory Comittee.*

ANDERSON, JOHN ROBERT (1983). «A spreading activation theory of memory», *Journal of Verbal Learning and Verbal Behavior*, **22**(3), 261–295.

ATKINS, BERYL y BETH LEVIN (1988). «Admitting impediments», en *Proceedings of the 4th Annual Conference of the UW Center for the New OED*, Oxford, UK.

BAR-HILLEL, Y. (1960). «Automatic translation of languages», *Advances in Computers.*

BAREISS, RAY (1990). *Exemplar-Based Knowledge Acquisition*, Academic Press.

BOGURAEV, BRANIMIR (1979). *Automatic Resolution of Linguistic Ambiguities*, Tesis Doctoral, Computer Laboratory, University of Cambridge.

BOOKMAN, LAWRENCE (1987). «A microfeature based scheme for modelling semantics», en *Proceedings of the 10th International Joint Conference on Artificial Intelligence, IJCAI´87*, págs. 611–614, Milán, Italy.

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.

BRISCOE, EDWARD (1991). «Lexical issues in natural language processing», en *Proceedings of the Symposium on Natural Language and speech*, págs. 39–68, Berlin.

BROWN, PETER F., STEPHEN A. DELLA PIETRA y VINCENT J. DELLA PIETRA (1991). «Word sense Disambiguation using statistical methods», en *Proceedings of 29th Annual Meeting of the Association for Computational Linguistics*, págs. 264–270.

BYRD, R. (1998). «Discovering Relationship among Word Senses», en *Proceedings of the 5th Annual Conference of the UW Centre for the New OED*, págs. 67–79.

CARMONA, J., S. CERVELL, L. MÀRQUEZ, M.A. MARTÍ, L. PADRÓ, H. RODRIGUEZ, R. PLACER, M. TAULÉ y J. TURMO (1998). «An Environment for Morphosyntactic Processing

of Unrestricted Spanish text», en *Proceedings of the First International Conference on Language Resources and Evaluation (LREC´98).*

CHAPMAN, ROBERT (1984). *Roget´s International Thesaurus, Fourth Edition*, Harper and Row.

CHEN, J. y J. CHANG (1998). «Topical Clustering of MRD Senses Based on Information Retrieval Techniques», *Computational Linguistics*, **24**(1), 61–95.

CHO, JEONG-MI y GIL C. KIM (1995). «Korean verb sense disambiguation using distributional information from corpora», en *Proceedings of the Natural Language Processing Pacific Rim Symposium*, págs. 691–696.

CHOUEKA, YAACOV y SERGE LUSIGNAN (1985). «Disambiguation by short contexts», *Computers and the Humanities*, **19**, 147–158.

CLARK, PETER y ROBIN BOSWELL (1991). «Rule induction with CN2: Some recent improvements», en *Proceedings of European Working Session on Learning*, págs. 151–163.

CLARK, PETER y TIM NIBLETT (1989). «The CN2 induction algorithm», *Machine Learning*, **3**(4), 261–283.

COTTRELL, GARRISON y STEVEN SMALL (1983). «A connectionist scheme for modelling word sense disambiguation», *Cognition and Brain Theory*, **6**, 89–120.

COWIE, JIM, JOE GUTHRIE y LOUISE GUTHRIE (1992). «Lexical disambiguation using simulated annealing», en *Proceedings of the 14th International Conference on Computational Linguistic, COLING´92*, págs. 359–365, Nantes, France.

DAELEMANS, W., J. ZAVREL, K. VAN DER SLOOT y A. VAN DER BOSCH (2001). «Timbl: Tilburg memory based learner, version 4.0, reference guide», *Technical report*, University of Antwerp.

DAGAN, IDO, FERNANDO PEREIRA y LILLIAN LEE (1994). «Word sense disambiguation using a second language monolingual corpus», *Computational Linguistics*, **20**(4), 563–596.

DAHLGREN, KATHLEEN (1988). *Naive Semantics for Natural Language Understanding*, Kluwer Academic Publishers, Norwell, MA.

ESCUDERO, G., L. MÁRQUEZ y G. RIGAU (2000a). «An Empirical Study of the Domain Dependence of Supervised Word Sense Disambiguation Systems», en *Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC'00)*, Hong Kong, China.

ESCUDERO, G., L. MÁRQUEZ y G. RIGAU (2000b). «Boosting Applied to Word Sense Disambiguation», en *Proceedings of the 11th European Conference on Machine Learning, ECML-2000*, Barcelona, Spain.

ESCUDERO, G., L. MÁRQUEZ y G. RIGAU (2000c). «Naive Bayes and Exemplar-Based approaches to Word Sense Disambiguation Revisited», en *Proceedings of the 14th European Conference on Artificial Intelligence, ECAI-2000*, Berlin, Germany.

FELLBAUM, CHRISTIANE (1998). *WordNet: An Electronic Lexical Database*, The MIT Press.

FERNÁNDEZ-AMORÓS, DAVID, JULIO GONZALO y FELISA VERDEJO (2001a). «The Role of Conceptual Relations in Word Sense Disambiguation», en *Proceedings 6th International Conference on Application of Natural Language to Information Systems (NLDB´2001).*, págs. 87–98, Madrid, Spain.

FERNÁNDEZ-AMORÓS, DAVID, JULIO GONZALO y FELISA VERDEJO (2001b). «The UNED systems at SENSEVAL-2», en ACL, editor, *Proceedings of the SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, págs. 75–78.

FLORIAN, R. y G.ÑGAI (2001). «Multidimensional transformational-based learning», en *Proceedings of the fifth Conference on Computational Natural Language Learning, CoNLL 2001*, págs. 1–8, Tolouse, France.

FUKUMOTO, FUMIYO y YOSHIMI SUZUKI (1996). «An automatic clustering of articles using dictionary definitions», en *Proceedings of 16th International Conference on Computational Linguistics*, págs. 406–411.

G. MILLER, T. RANDEE, C. LEACOCK y R. BUNKER (1993). «A Semantic Concordance», en *Proceeding of 3rd DARPA*

Workshop on Human Language Tecnology, págs. 303–308, Plainsboro, New Jersey.

GALE, W. y K. CHURCH (1991). «A program for aligning sentences in bilingual corpora», en *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistic*, págs. 177–184.

GALE, W., K. CHURCH y D. YAROWSKY (1992a). «Estimating upper and lower bounds on the performance or word-sense disambiguation programs», en *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistic.*

GALE, WILLIAM, KENNETH CHURCH y DAVID YAROWSKY (1992b). «A method for disambiguating word senses in a large corpus», *Computers and the Humanities*, **26**, 415–439.

GALE, WILLIAM, KENNETH CHURCH y DAVID YAROWSKY (1992c). «One sense per discourse», en Morgan Kaufmann, editor, *Proceedings of the speech and Natural Language Workshop*, págs. 233–237, San Francisco, USA.

GONZALO, J., A. PEÑAS y F. VERDEJO (1999). «Lexical ambiguity and Information Retrieval Revisited», en *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99)*, MaryLand.

GONZALO, J., F. VERDEJO, I. CHUGUR y J. CIGARRÁN (1998). «Indexing with WordNet synsets can improve text retrieval», en *Proceedings of the ACL/COLING Workshop on Usage of WordNet for Natural Language Processing* .

GOUGENHEIM, GEORGES y RENÉ MICHÉA (1961). «Sur la determination du sens d´un mot au moyen du contexte», *La Traduction automatique*, **2**(1), 16–17.

GRISHMAN, RALPH, CATHERINE MACLEOD y ADAM MEYERS (1994). «COMLEX syntax: Building a computational lexicon. », en *Proceedings of the 15th International Conference on Computational Linguistic, COLING´94*, págs. 268–272, Kyoto, Japan.

GUTHRIE, JOE, LOUISE GUTHRIE, YORICK WILKS y HOMA AIDINEJAD (1991). «Subject-dependent co-occurrence and word sense disambiguacion», en *Proceedings of the 29th Annual Mee-*

ting of the Association for Computational Linguistics, págs. 146–152, Berkeley, CA.

HARPER, KENNETH (1957a). «Contextual analysis», *Mechanical Translation*, **4**(3), 70–75.

HARPER, KENNETH (1957b). «Semantic ambiguity», *Mechanical Translation*, **4**(3), 68–69.

HAYES, PHILIP (1977a). «On semantics nets, frames and associations», en *Proceedings of 5th International Joint Conference on Artificial Intelligence*, págs. 99–107, Cambridge, MA.

HAYES, PHILIP (1977b). *Some Association-based Techniques for Lexical Disambiguation by Machine*, Tesis Doctoral, Ecole Polytechnique Fédérale de Lausanne.

HAYES, PHILIP (1978). «Mapping input into schemas», *inf. téc.*, 29, Department of Computer Science, University of Rochester.

HIRST, GRAEME (1987). *Semantic Interpretation and the Resolution of Ambiguity*, Cambridge University Press.

HOSTE, VERÓNIQUE, ANNE KOOL y WALTER DAELEMANS (2001). «Classifier optimization and combination in the English all words task», en ACL, editor, *Proceedings of the SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, págs. 84–86.

IDE, N. y J. VÉRONIS (1998). «Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art», *Computational Linguistics*, **24**(1), 1–40.

JUSTESON, JOHN y SLAVA KATZ (1995). «Principled disambiguation: Discriminating adjective senses with modified nouns», *Computational Linguistics*, **21**(1), 1–27.

KAPLAN, ABRAHAM (1950). «An experimental study of ambiguity and context», *Mechanical Translation*, **2**(2), 39–46.

KAWAMOTO, ALAN (1988). *Distributed representations of ambiguous words and their resolution in a connectionist network*, Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology and Artificial Intelligence, Morgan Kaufman, Sam Mateo, CA, steven small and garrison cottrell and michael tanenhaus ed.

KELLY, EDWARD y PHILIP STONE (1975). *Computer Recognition of English Word Senses*, North-Holland, Amsterdam.

KER, SUE y JASON CHANG (1997). «A class-based approach to word alignment», *Computational Linguistics*, **23**(2), 313–343.

KILGARRIFF, A. (1998). «Gold standard for evaluating word sense disambiguation programs», *Computer Speech and Language, Special Issue on evaluation*, **12**(3).

KILGARRIFF, A. y J ROSENZWEIG (2000). «Framework and results for english SENSEVAL», *Computers and the Humanities*, **34**(1-2).

KILGARRIFF, ADAM (1994). «The myth of completeness and some problems with consistency (the role of frequency in deciding what goes in the dictionary)», en *Proceedings of the 6th International Congress on Lexicography, EURALEX´94*, págs. 101–106, Amsterdam, Holland.

KILGARRIFF, ADAM (1997). «What is word sense disambiguation good for?», en *Proceedings of the Natural Language Processing Pacific Rim Symposium*, págs. 209–214.

KILGARRIFF, ADAM y MARTHA PALMER (2000). «Introduction to the Special Issue on SENSEVAL.», *Computers and the Humanities*, **34**(1-2), 1–13.

KOLODNER, JANET (1993). *Case-Based Reasoning*, Morgan Kaufmann.

KOUTSOUDAS, ANDREAS y R. KORFHAGE (1956). «MT and the problem of the multiple meaning», *Mechanical Translation*, **2**(2), 46–51.

KROVETS, R. (1997). «Homonymy and polysemy in information retrival», en *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistic and 8th Conference of the European Chapter of the Association for Computational Linguistic*, págs. 72–79, Madrid, Spain.

KROVETS, R. y W. BRUCE CROFT (1992). «Lexical ambiguity and information retrieval», *ACM*, **10**(2), 115–141.

KROVETZ, ROBERT y WILLIAM B. CROFT (1989). «Word Sense Disambiguation using machine-readable dictionaries», en *Proceedings of the 12th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval SIGIR-89*, págs. 127–136, Cambridge, MA.

LEACOCK, CLAUDIA, MARTIN CHODOROW y GEORGE A. MILLER (1998). «Using Corpus Statistics and WordNet Relations for Sense Identification», *Computational Linguistics*, **24**(1), 147–165.

LEACOCK, CLAUDIA, GEOFFREY TOWELL y ELLEN VOORHEES (1993). «Corpus based statistic sense resolution», en *Proceedings of the ARPA Human Language Technology Workshop*, págs. 260–265, San Francisco, Morgan Kaufman.

LENAT, DOUGLAS (1995). «Cyc: A Large-Scale Investment in Knowledge Infrastructure», *Communications of the ACM*, **38**(11).

LESK, MICHAEL (1986). «Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone», en *Proceedings of the 1986 SIGDOC Conference, Association for Computing Machinery*, págs. 24–26, Toronto, Canada.

LI, H. y N. ABE (1995). «Generalizing Case Frames Using a Thesaurus and the MDL Principle», en *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP´95)*, págs. 239–248.

LI, XIAOBIN, STAN SZPAKOWICZ y STAN MATWIN (1995). «A WordNet-based algorithm for word sense disambiguation», en *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, págs. 1368–1374.

LIDDY, ELISABETH y WOOJIN PAIK (1993). «Statistically-guided word sense disambiguation», en *Proceedings of the AAAI Fall Symposium Series*, págs. 98–107.

LIN, DEKANG (1997). «Using syntactic dependency as local context to resolve word sense ambiguity», en *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistic and the 8th Conference of the European Chapter of the Association for Computational Linguistics*, págs. 64–71.

LITOWSKI, KENNETH (1997). «Desiderata for tagging with WordNet synsets or MCAA categories», en *Proceedings of the ACL-SIGLEX Workshop "Tagging Text with Lexical Semantics: Why What, and How?*, págs. 12–17, Washington, USA.

LYTINEN, STEVEN (1986). «Dynamically combining syntax and semantics in natural language processing», en *Proceedings of AAAI-86*, págs. 574–578.

MAGNINI, B. y G. CAVAGLIA (2000). «Integrating subject field codes into WordNet», en *Proceedings of Third International Conference on Language Resources and Evaluation (LREC-2000)*.

MAGNINI, BERNARDO y C. STRAPPARAVA (2000). «Experiments in Word Domain Disambiguation for Parallel Texts», en *Proceedings of the ACL Workshop on Word Senses and Multilinguality*, Hong Kong, China.

MANNING, C. D. y H. SCHÜTZE (1999). *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, Massachusetts.

MARTÍ, M.A., H. RODRÍGUEZ y J. SERRANO (1998). «Declaración de categorías morfosintácticas», Proyecto ITEM. Doc. nº2, `http://sensei.ieec.uned.es/item`.

MARTÍNEZ, D. y E. AGIRRE (2000). «One sense per collocation and genre/topic variations», en *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Hong Kong.

MARUYAMA, HIROSHI (1990). «Structural disambiguation with constraint propagation», en *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistic*, págs. 31–38.

MASTERMAN, M. (1957). «The thesaurus in syntax and semantics», *Connectionist, statistical and simbolic approaches to learning for Natural Language Processing*, 4, 1–2.

MASTERMAN, M. (1962). «Semantic message detection for machine translation, using an interlingua», en *Proceedings of International Conference on Machine Translation of Languages and Applied Language Analysis*, págs. 437–475.

MCCARTHY, D. (2001). *Lexical Acqusition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences. Gzipped postscript version available*, Tesis Doctoral, Cognitive and Computing Sciences. University of Sussex.

McRoy, Susan (1992). «Using multiple knowledge sources for word sense discrimination», *Computational Linguistics*, **18**(1), 1–30.

Mihalcea, Rada (2001). «Gencor: a large semantically tagged corpus», en *preparation.*

Mihalcea, Rada y Dan Moldovan (1998). «Word Sense Disambiguation based on Semantic Density», en *Proceedings of COLING-ACL'98 Workshop on Usage of WordNet in Natural Language Processing Systems*, págs. 16–22, Montreal, Canada.

Mihalcea, Rada y Dan Moldovan (1999a). «A Method for word sense disambiguation of unrestricted text», en *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistic*, págs. 152–158, Maryland, Usa.

Mihalcea, Rada y Dan Moldovan (1999b). «An automatic method for generating sense tagged corpora», en *Proceedings of AAAI-99*, Orlando, Usa.

Mihalcea, Rada y Dan Moldovan (2000). «An Iterative Approach to Word Sense Disambiguation», en *Proceedings of Flairs 2000*, págs. 219–223, Orlando, FL.

Miller, G., R. Beckwith, C. Fellbaum, D. Gross y K. Miller (1990). «WordNet: An on-line lexical database», *International journal of lexicography*, **3**(4), 235–244.

Montoyo, Andrés (2000). «Método basado en Marcas de Especificidad para WSD», *Procesamiento del Lenguaje Natural*, **26**, 53–59.

Montoyo, Andrés y Manuel Palomar (2000a). «Word Sense Disambiguation with Specification Marks in Unrestricted Texts.», en *Proceedings of 11th International Workshop on Database and Expert Systems Applications (DEXA 2000). 11th International Workshop on Database and Expert Systems Applications*, págs. 103–107, IEEE Computer Society, Greenwich, London, UK.

Montoyo, Andrés y Manuel Palomar (2000b). «WSD Algorithm Applied to a NLP System », en Mokrane Bouzeghoub, Zoubida Kedad y Elisabeth Métais, editores, *Proceedings of 5th International conference on Applications of Natural Language to Information Systems (NLDB-2000). Natural Language Pro-*

cessing and Information Systems, Lecture Notes in Computer Science, págs. 54–65, Springer-Verlag, Versailles, France.

MONTOYO, ANDRÉS y MANUEL PALOMAR (2001). «Specification Marks for Word Sense Disambiguation: New Development», en A. Gelbukh, editor, *Proceedings of 2nd International conference on Intelligent Text Processing and Computational Linguistics (CICLing-2001). Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, págs. 182–191, Springer-Verlag, Mexico City.

MONTOYO, ANDRÉS, MANUEL PALOMAR y GERMAN RIGAU (2001). «WordNet Enrichment with Classification Systems.», en *Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customisations Workshop. (NAACL-01) The Second Meeting of the North American Chapter of the Association for Computational Linguistics*, págs. 101–106, Carnegie Mellon University. Pittsburgh, PA, USA.

MOONEY, RAYMOND (1996). «Comparative experiments on disambiguating word sense: An illustration of the role of bias in machine learning», en *Proceedings of the 1st Conference on Empirical Methods in Natural Language Processing*, págs. 82–91.

MÁRQUEZ, L. (1999). *Part-of-Speech Tagging: A Machine Learning Approach Based on Decision Trees*, Tesis Doctoral, Universidad Politécnica de Cataluña, Barcelona.

NAGAO, KATASHI (1994). «A preferential constraint satisfaction technique for natural language analysis», *IEICE Transactions on Information and Systems*, **E77-D**(2), 161–170.

NG, HWEE TOU (1997a). «Exemplar-based word sense disambiguation: Some recent improvements», en *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*, págs. 208–213.

NG, HWEE TOU (1997b). «Getting Serious about Word Sense Disambiguation», en *Proceedings of the SIGLEX Workshop "Tagging Text with Lexical Semantics: Why, What and How?*

NG, HWEE TOU y HIANG BENG LEE (1996). «Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach», en *Proceedings of the 34th Annual*

Meeting of the Association for Computational Linguistic, págs. 40–47, University of California, Santa Cruz, CA.

OKUMURA, MANABU y H. TANAKA (1990). «Towards Incremental Disambiguation with a Generalized Discrimination Network», en *Proceedings of 8th National Conference on Artificial Intelligence*, págs. 990–995.

PADRÓ, L. (1998). *A Hybrid environment for Syntax-Semantic tagging*, Tesis Doctoral, Universidad Politécnica de Cataluña, Barcelona.

PEDERSEN, TED y REBECCA BRUCE (1997a). «A new supervised learning algorithm for word sense disambiguation», en *Proceedings of AAAI/IAAA-97*, págs. 604–609.

PEDERSEN, TED y REBECCA BRUCE (1997b). «Distinguishing word senses in untagged text», en *Proceedings of the 2th Conference on Empirical Methods in Natural Language Processing*, págs. 197–207.

PEDERSEN, TED, REBECCA BRUCE y JANYCE WIEBE (1997). «Sequential model selection for word sense disambiguation», en *Proceedings of the 5th Conference on Applied Natural Language Processing*, págs. 388–395.

PLA, F. (2000). *Etiquetado Léxico y Análisis Sintáctico Superficial basado en Modelos Estadísticos*, Tesis doctoral, Departamento de Sistemas Informáticos y Computación. Universidad de Politécnica de Valencia.

PLA, F., A. MOLINA y N. PRIETO (2000). «Tagging and Chunking with Bigrams», en *Proceeding of 19th International Conference on Computational Linguistics (COLING'2000)*, Saarbrucken, Germany.

QUILLIAN, M. ROSS (1968). *Semantic memory*, págs. 227–270, MIT Press, M. Minsky ed.

QUILLIAN, M. ROSS (1969). «The teachable language comprehender: A simulation program and theory of language», *Communications of the ACM*, **12**(8), 459–476.

QUINLAN, ROSS (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann.

RADA, R., H. MILI, E. BICKNELL y M. BLETTNER (1989). «Development an Application of a Metric on Semantic Nets», *IEEE Transactions on Systems, Man and Cybernetics*, **19**(1), 17–30.

RAVIN, YAEL (1990). «Disambiguating and interpreting verb definitions», en *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistic*, págs. 260–267.

RESNIK, P. y D. YAROWSKY (1999). «Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation», *Natural Language Engineering*, **5**(2), 113–134.

RESNIK, PHILIP (1993a). *Selection and Information: A Class-Based Approach to Lexical Relationships*, Tesis Doctoral, University of Pennsylvania.

RESNIK, PHILIP (1993b). *Selection and Information: A Class-based Approach to Lexical Relationships*, Tesis Doctoral, University of Pennsylvania.

RESNIK, PHILIP (1993c). «Semantic classes and syntactic ambiguity», en *Proceedings of the ARPA Workshop on Human Human Language Technology*, págs. 278–283.

RESNIK, PHILIP (1995a). «Disambiguating noun groupings with respect to WordNet senses», en *Proceedings of the Third Workshop on Very Large Corpora*, págs. 54–68, Cambridge, MA.

RESNIK, PHILIP (1995b). «Using information content to evaluate semantic similarity in a taxonomy», en *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, págs. 448–453.

RESNIK, PHILIP y DAVID YAROWSKY (1997). *A perspective on word sense disambiguation methods and their evaluation*, Tagging Text with Lexical Semantics: Why, What and How?, ACL SIGLEX, m. light ed.

RIBAS, F. (1995a). *On Acquiring Appropiate Selectional Restrictions from Corpora Using a Semantic Taxonomy*, Tesis Doctoral, Universitat Politècnica de Catalunya, Barcelona.

RIBAS, F. (1995b). «On learning more appropiate selectional restrictions», en *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, Dublin, Ireland.

RICHARDSON, S. (1997). *Determining Similarity and Inferring Relations in a Lexical Knowledge Base*, Tesis Doctoral, The City University of New York.

RICHENS, R. H. (1996). «Interlingual machine translation», *Connectionist, statistical and simbolic approaches to learning for Natural Language Processing*, **1**(3), 144–147.

RIGAU, G. (1994). «An Experiment on Automatic Semantic Tagging of Dictionary Senses», en *Proceedings of the International Workshop the Future of the Dictionary.*

RIGAU, G., H. RODRÍGUEZ y E. AGIRRE (1998). «Building Accurate Semantic Taxonomies from Monolingual MRDs», en *Proceedings of the 17th International Conference on Computational Linguistic and 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL´98*, Montreal, Canada.

RIGAU, GERMAN, ENEKO AGIRRE y JORDI ATSERIAS (1997). «Combining unsupervised lexical knowledge methods for word sense disambiguation», en *Proceedings of joint 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics ACL/EACL'97*, Madrid, Spain.

RIVEST, RONALD (1987). «Learning decision lists», *Machine Learning*, **2**, 229–246.

ROBERTS, D. (1973). *The existencial Graphs of Charles S. Pierce*, Mouton, The Hague.

SALTON, G. (1989). *Automatic Text Processing: the transformation, analysis and retrieval of information by computer*, Addison Wesley.

SANDERSON, M. (1994). «Word sense disambiguation and information retrieval», en *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, págs. 142–151.

SCHMID, HELMUT (1994). «Probabilistic part-of-speech tagging using decision trees», en *Proceedings International Conference on New Methods in Language Processing.*, págs. 44–49, Manchester, UK.

SCHÜTZE, H. (1992). «Dimensions of meaning», en *Proceedings of Supercomputing*, págs. 787–796.

SCHÜTZE, H. (1995). «Information retrieval based on word senses», en *Proceedings of 4th Annual Symposium on Document Analysis and Information Retrieval*, págs. 161–175.

SCHÜTZE, H. (1998). «Automatic word sense discrimination», *Computational Linguistics*, **24**(1), 97–123.

SELZ, O. (1922). *Zue Psychologie des produktive Denkens un des Irrtums*, Friedrich Cohen, Bon.

SLATOR, B. M. y Y. WILKS (1987). «Towards semantic structures from dictionary entries», en *Proceedings of the 2nd annual rocky mountain conference on Artificial Intelligence*, págs. 85–96.

SLATOR, BRIAN (1992). «Sense and preference», *Computer and Mathematics with Applications*, **23**(6/9), 391–402.

SMALL, S. (1983). «Parsing as cooperative distributed inference: Understanding through memory interactions», en M. King, editor, *Parsing Natural Language*, págs. 247–276, Academic Press, London.

SMALL, STEVEN, W. COTTRELL GARRISON y MICHAEL K. TANENHAUS (1988). *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology and Artificial Intelligence*, Morgan Kaufmann Publishers Inc, San Mateo, CA.

SMALL, STEVEN y CHARLES RIEGER (1982). *Parsing and comprehending with word experts (a theory and its realization)*, págs. 89–147, Lawrence Erlbaum and associates, Hillsdale, NJ, Wendy Lenhert and Martin Ringle ed.

STETINA, J., S. KUROHASHI y M.ÑAGAO (1998). «General word sense disambiguation method based on full sentencial context», en *Proceedings of Usage of WordNet in Natural Language Processing. COLING-ACL Workshop*, Montreal, Canada.

STONE, PHILIP (1969). *Improved quality of content analysis categories: Computerized-disambiguation rules for high-frequency English words*, págs. 199–221, John Wiley and Sons, George Gerbner and Ole Holsti and Klaus Krippendorf and William Paisley and Philip Stone ed.

STONE, PHILIP, DEXTER DUNPHY, MARSHALL SMITH y DANIEL OGILVIE (1966). *The General Inquier: A Computer Approach to Content Analysis*, MIT Press, Cambridge, MA.

SUÁREZ, A. y M. PALOMAR (2002). «Feature Selection Analysis for Maximum Entropy-based WSD», en *Proceedings of Third International conference on Intelligent Text Processing and Computational Linguistics (CICLing-2001). Computational Linguistics and Intelligent Text Processing*, Mexico City.

SUÁREZ, A. y A. MONTOYO (2001). «Estudio de cooperación de métodos de desambiguación léxica: marcas de especificidad vs. máxima entropía», *Revista Procesamiento del Lenguaje Natural*, **27**(1), 207–214.

SUSSNA, MICHAEL (1993). «Word sense disamiguation for free-text indexing using a massive semantic network. », en *Proceedings of the Second International Conference on Information and Knowledge Base Management, CIKM´93*, págs. 67–74, Arlington, VA.

TANAKA, HIDEKI (1994). «Verbal case frame acquisition from bilingual corpus: Gradual knowledge acquisition», en *Proceedings of the 15th International Conference on Computational Linguistics*, págs. 727–731.

UREÑA, ALFONSO (1998). *Resolución de la ambigüedad léxica en tareas de clasificación automática de documentos*, Tesis Doctoral, Universidad de Jaen.

VERDEJO, FELISA (1994). «Comprensión del lenguaje natural: avances, aplicaciones y tendencias en procesamiento del lenguaje natural: fundamentos y aplicaciones», en *In Documentación del curso de verano de 1994 de la Universidad Nacional de Educación a Distancia*, Ávila.

VOORHEES, ELLEN (1993). «Using WordNet to disambiguate word senses for text retrieval», en *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, págs. 171–180, Pittsburgh, PA.

VOSSEN, P. (1998). «The Restructured Core WordNets in EutoWordNet: Subset1.», en *Deliverable D014, D015, WP3, WP4*, EuroWordNet LE2-4003.

VOTILAINEN, A. (1988). «NPTool, a Detector of English Noun Prhase», en *Proceedings of the Workshop on Very Large Corpora, ACL.*

VOTILAINEN, A. y T. JÄRVINEN (1995). «Specifying a Shallow Grammatical Representation for Parsing Purposes», en *Proceedings of the 7th European Chapter of the Association for Computational Linguistics (EACL)*, Dublin, Ireland.

WALKER, DONALD (1987). *Knowledge resource tools for accessing large text files*, Machine Translation: Theorical and methodological issues, Cambridge: Cambridge University Press, ralph grishman and richard kittredge ed.

WALTZ, DAVID y JORDAN POLLACK (1985). «Massively parallel parsing: A strongly interactive model of natural language interpretation», *Cognitive Science*, **9**, 51–74.

WEAVER, WARREN (1955). «Translation», *Machine Translation of Languages*, págs. 15–23.

WEISS, S. y C. KULIKOWSKI (1991). *Computer Systems That Learn*, Morgan Kaufmann.

WILKS, Y., B. SLATOR y L. GUTHRIE (1996). *Electric words: Dictionaries, Computers and Meanings*, MIT Press, Cambridge, MA.

WILKS, Y. y M. STEVENSON (1996). «The grammar of sense: Is word sense tagging much more than part-of-speech tagging?», *inf. téc.*, CS-96-05, University of Sheffield. UK.

WILKS, Y. y M. STEVENSON (1997). «Combining independent knowledge sources for word sense disambiguation», en *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP).*

WILKS, YORICK (1968). «On-line semantic analysis of English texts», *Mechanical Translation*, **11**(3-4), 59–72.

WILKS, YORICK (1969). «Getting meaning into the machine», *New Society*, **361**, 315–317.

WILKS, YORICK (1973). *An artificial intelligence approach to machine translation*, págs. 114–151, W. H. Freeman, San Francisco, Roger Schank and Kenneth Colby ed.

WILKS, YORICK (1975a). «A preferential pattern-seeking semantics for natural language inference», *Artificial Intelligence*, **6**, 53–74.

WILKS, YORICK (1975b). «An intelligent analyzer and understander of English», *Communications of the ACM*, **18**(5), 264–274.

WILKS, YORICK (1975c). *Preference semantics*, págs. 329–348, Cambridge University Press, E. L. Keenan III ed.

WILKS, YORICK (1975d). «Primitives and words», en *Proceedings of the Interdisciplinary Workshop on Theorical Issues in Natural Language Processing*, págs. 42–45, Cambridge, MA.

WILKS, YORICK, DAN FASS, CHENG-MING GUO, JAMES E. MACDONALD, TONY PLATE y BRIAN SLATOR (1990). «Providing machine tractable dictionary tools», en James Pustejovsky, editor, *Semantics and the Lexicon*, MIT Press, Cambridge, MA.

YAROWSKY, D. (1993). «One sense per collocation», en *Proceedings of the DARPA Workshop on Human Language Technology*, págs. 266–271, Princenton, NJ.

YAROWSKY, D. (2000). «Hierarchical decision lists for word sense disambiguation», *Computers and the Humanities*, **34**, 1–2.

YAROWSKY, D. y R. WICENTOWSKI (2000). «Minimally supervised morphological analysis by multimodal alignment», en *Proceedings of Proceedings of the 38th Annual Meeting of the Association for Computational Linguistic ACL-2000*, págs. 207–216, Hong Kong.

YAROWSKY, DAVID (1992). «Word sense disambiguation using statistical models of Roget´s categories trained on large corpora», en *Proceedings of the 14th International Conference on Computational Linguistic, COLING´92*, págs. 454–460, Nantes, France.

YAROWSKY, DAVID (1994a). «A comparison of corpus-based techniques for restoring accents in Spanish and French text», en *Proceedings of the 2th Annual Workshop on Very Large Text Corpora*, págs. 19–32.

YAROWSKY, DAVID (1994b). «Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and

French», en *Proceedings of the 32th Annual Meeting of the Association for Computational Linguistic*, págs. 88–95.

YAROWSKY, DAVID (1995). «Unsupervised word sense disambiguation rivaling supervised methods», en *Proceedings of the 33th Annual Meeting of the Association for Computational Linguistic*, págs. 189–196.

YNGVE, VICTOR (1955). «Syntax and the problem of multiple meaning», *Machine Translation of Languages*, págs. 208–226.

UNIVERSIDAD DE ALICANTE

Comisión de Doctorado

Reunido el Tribunal que suscribe en el día de la fecha acordó otorgar, por UNANIMIDAD la Tesis Doctoral de Don/Dña. J. Andrés Montoyo Guijarro la calificación de SOBRESALIENTE CUM LAUDE

Alicante 21 de JUNIO de 2002

El Presidente,

LLUÍS PADRÓ

El Secretario,

P. MARTÍNEZ BARCO

UNIVERSIDAD DE ALICANTE

Comisión de Doctorado

La presente Tesis de D. J. Andrés Montoyo Guijarro ha sido registrada al Folio con el n.º 858-041 del registro de entrada correspondiente.

Alicante 26 de junio de 2002

El Encargado del Registro,

Volver al índice/Tornar a l'índex