Resolución de la Ambigüedad Léxica en Tareas de Clasificación Automática de Documentos

L. Alfonso Ureña López

Índice General

1 Introducción **9**

2 Análisis del contenido **15**
2.1 Introducción . . . 15
2.2 Tareas . . . 16
2.2.1 Clasificación de documentos . . . 16
2.2.2 Resolución de la ambigüedad léxica . . . 20
2.2.3 Traducción automática . . . 21
2.2.4 Generación automática de resúmenes . . . 22
2.3 Técnicas de indexación automática . . . 23
2.3.1 El modelo del espacio vectorial . . . 23
2.3.2 Pesos de términos . . . 25
2.3.3 Listas de parada . . . 26
2.3.4 Extracción de raíces . . . 27
2.3.5 Realimentación . . . 28
2.4 Resumen y conclusiones . . . 28

3 Recursos lingüísticos **31**
3.1 Introducción . . . 31
3.2 Corpora de textos . . . 32
3.2.1 Características de un corpus . . . 33
3.2.2 Corpora anotados o etiquetados . . . 34
3.3 Lexicón . . . 42
3.3.1 Bases de datos léxicas . . . 43
3.3.2 Diccionarios electrónicos . . . 48
3.3.3 Thesaurus . . . 50
3.4 Resumen y conclusiones . . . 51

4 Resolución de la ambigüedad léxica **53**
4.1 Introducción . . . 53
4.2 Descripción de la tarea y terminología . . . 54

4.2.1 Contexto . . . 54
4.2.2 Sentidos o usos . . . 54
4.2.3 Granularidad de sentidos . . . 55
4.3 Investigación en desambiguación . . . 56
4.3.1 Desambiguación basada en reglas generadas manualmente . . . 56
4.3.2 Métodos basados en corpora de textos . . . 58
4.3.3 Métodos basados en el conocimiento . . . 61
4.4 Efectividad del proceso de desambiguación . . . 66
4.4.1 Evaluación de la desambiguación . . . 70
4.5 Sistema WSD . . . 72
4.5.1 Idea básica . . . 72
4.5.2 Metodología . . . 73
4.5.3 Algoritmos de aprendizaje . . . 75
4.6 Enfoques WSD . . . 77
4.6.1 Desambiguador basado en corpus . . . 77
4.6.2 Desambiguador basado en WordNet . . . 80
4.6.3 Desambiguador basado en la integración de recursos lingüísticos . . . 82
4.6.4 Estructura del desambiguador . . . 84
4.7 Descripción del entorno experimental . . . 86
4.7.1 Recursos empleados . . . 86
4.7.2 Tamaño de la ventana contextual . . . 87
4.7.3 Resultados experimentales de la evaluación directa e interpretación . . . 89
4.8 Resumen y conclusiones . . . 98

5 Aplicación de WSD a tareas de clasificación de documentos 101
5.1 Introducción . . . 101
5.2 Aplicando la desambiguación a la Recuperación de Información . . . 102
5.2.1 Procesamiento de la consulta . . . 103
5.2.2 La tarea de recuperación con realimentación por relevancia . . . 104
5.2.3 Utilización de WordNet en la recuperación . . . 106
5.2.4 Entorno de evaluación . . . 109
5.3 Integrando recursos lingüísticos en TC por medio de WSD . . . 121
5.3.1 Uso de WordNet en la categorización . . . 122
5.3.2 Uso de la desambiguación para la categorización basada en WordNet . . . 125
5.3.3 Experimentos centrados en la efectividad . . . 125
5.4 Resumen y conclusiones . . . 130

6 Conclusiones 131
6.1 Principales aportaciones . . . 131
6.2 Futuros desarrollos . . . 132

A SemCor **135**
A.1 Estadísticas 135
A.2 Estructura de los documentos 136
A.3 Descripción de los elementos SGML 136
A.4 Etiquetas sintácticas 137

B WordNet **139**
B.1 Estructura de la base de datos de WordNet 139

C Detalles adicionales de los experimentos **143**
C.1 Experimentos WSD basados en SemCor 143
C.2 Experimentos WSD basados en WordNet 145

D Lista de Topics y expresiones textuales asociadas **153**

Bibliografía **158**

Índice de Figuras

2.1 Recuperación de textos 16
2.2 Interfaz del sistema infoseek 17
2.3 Categorización de textos 18
2.4 Resolución de la ambigüedad 20
2.5 Sistema SYSTRAN 21
2.6 Histograma de la frecuencia de las palabras mostrando la ley de Zipf 26

3.1 Fragmento del corpus LEXESP 38
3.2 Corpus paralelo de la Biblia Políglota 39
3.3 Fragmento de un texto de SEMCOR 40
3.4 Documento 18.753 de Reuters-21.578 42
3.5 Estructura y relaciones de WORDNET 46
3.6 Muestras de información asociada en WORDNET 47
3.7 Muestra del thesaurus Roget con información relativa al término "car" 49
3.8 Fragmentos de thesaurus 50
3.9 Ejemplo de lista de asociación de términos 51

4.1 Partición de términos 68
4.2 Vectores representando significados y consultas 75
4.3 El proceso de desambiguación 77
4.4 Aquitectura del enfoque WSD basado en el recurso lingüístico SEMCOR 78
4.5 Vectores representando *ventanas contextuales* 78
4.6 Ejemplo de Ventana Contextual (fragmento extraído del documento *br-c02* de SEMCOR) 79
4.7 Fragmento de la relación jerárquica en WORDNET 81
4.8 Arquitectura del enfoque WSD basado en base de datos léxica (WORDNET) . . . 81
4.9 Ejemplo de información de sinonimia e hiperonimia extraída de WORDNET para el término *bank* 82
4.10 Arquitectura basada en la integración de recursos lingüísticos, corpus de entrenamiento y base de datos léxica 83
4.11 DFD que describe el enfoque de desambiguación basado en SEMCOR 84

4.12 DFD que describen el enfoque de desambiguación basado en WordNet 84
4.13 DFD que describe el enfoque de desambiguación basado en la integración de recursos lingüísticos 85
4.14 *Microaveraging* y *macroaveraging* para una ventana contextual con un tamaño en el intervalo [10,60] 88
4.15 *microaveraging* y *macroaveraging* para una ventana contextual de tamaño "párrafo" 89
4.16 Ejemplo de Ventanas Contextuales construidas a partir de SemCor, para el término "bank" con los significados #1, #5 y #7 91
4.17 Ejemplo de Ventanas Contextuales construidas a partir de todas las relaciones de WordNet, para el término "bank" con los significados #1, #5 y #7 92
4.18 Distribución de sentidos en la colección SemCor 1.6 (escala logarítmica) 93
4.19 Relación entre *precision* (*microaveraging* y *macroaveraging*) y *recall* basada en las relaciones de WordNet 96

5.1 El sistema WordNet y aplicaciones 102
5.2 Proceso de recuperación de información con expansión de la consulta 108
5.3 Expansión total para los términos de la consulta número 10 (TREC) con información de WordNet 112
5.4 Consulta TREC (*Topic* número 10) 113
5.5 Fragmento de un documento del Wall Street Journal (documento: WSJ900416-0096) 114
5.6 Efectividad para los diferentes tipos de expansión con *feedback* 116
5.7 Términos que aparecen cerca de una categoría 121
5.8 Proceso de integración de recursos lingüísticos 123
5.9 Efectividad para los diferentes tipos de categorización 128
5.10 Consulta en WordNet de la categoría "inventories" 129

B.1 Muestra del fichero "noun.idx" 141
B.2 Muestra del fichero "noun.dat" 142

Índice de Tablas

2.1 Primeros términos de la lista de parada utilizada en SMART 27

3.1 Contenido de SEMCOR 1.6 41
3.2 Clasificación de categorías Reuters-21578 41

4.1 Porcentaje de ocurrencias de SEMCOR 1.6 90
4.2 Resultados de los experimentos para el enfoque basado en corpus de entrenamiento 94
4.3 Evaluación realizada entrenando con el Brown2 y evaluando con el Brown1 94
4.4 *precision* y *recall* obtenidos por las distintas relaciones léxicas y semánticas del enfoque basado en WORDNET en la evaluación del Brown1 95
4.5 Ejemplo de combinación de relaciones 97
4.6 Resultados obtenidos entrenando con el *Brown2* e integrando el entrenamiento con información de sinonimia de WORDNET 98
4.7 *precision* y *recall* obtenidos por la combinación de recursos lingüísticos en la evaluación de la colección de prueba *Brown1* de SEMCOR 99

5.1 Número de términos contenidos en el corpus WSJ y en las consultas TREC . . . 107
5.2 Características de la colección de prueba utilizada en la evaluación de la recuperación 110
5.3 Dominios de las consultas utilizadas en los experimentos 110
5.4 *Precision media* y porcentaje de cambio en los 11 niveles estándar y 3 intermedios (0.2, 0.5, 0.8) de *recall* 117
5.5 Evaluación en 11 niveles de *Recall*, obtenida con realimentación por relevancia para los procesos de expansión 118
5.6 *Precision* y *Recall* en 5, 10, 15 y 30 documentos 119
5.7 Valores exactos de *precision* y *recall* para consultas con 1 y 2 términos 119
5.8 Efectividad de la estrategia de Voorhees [1994] cuando se expande la consulta con los *synsets* seleccionados automáticamente 120
5.9 Estadísticas de la colección de documentos Reuters-21578 126
5.10 *Precision* en 11 niveles de *recall* 127
5.11 f_1 calculada por medio de *macro* y *microaveraging*. 127

A.1 Estadísticas de SEMCOR 135

A.2 Valores de `cm` 137
A.3 Etiquetas sintácticas de SemCor 138

B.1 Ficheros de la base de datos léxica WordNet (v. 1.6) 140

C.1 Resultados correspondientes a la evaluación de los 50 primeros documentos del Brown1 de SemCor 144
C.2 Experimentos correspondientes a los primeros 50 documentos del Brown1 mediante la relación *sinonimia* de WordNet 146
C.3 Experimentos correspondientes a los primeros 50 documentos del Brown1 mediante la relación *hiponimia* de WordNet 147
C.4 Experimentos correspondientes a los primeros 50 documentos del Brown1 mediante la relación *hiperonimia* de WordNet 148
C.5 Experimentos correspondientes a los primeros 50 documentos del Brown1 mediante la relación *meronimia* de WordNet 149
C.6 Experimentos correspondientes a los primeros 50 documentos del Brown1 mediante la relación *holonimia* de WordNet 150
C.7 Experimentos correspondientes a los primeros 50 documentos del Brown1 mediante la relación *antonimia* de WordNet 151

D.1 Lista de *Topics* utilizados en la categorización (I) 154
D.2 Lista de *Topics* utilizados en la categorización (y II) 155
D.3 Lista de *Topics* utilizados en la desambiguación (I) 156
D.4 Lista de *Topics* utilizados en la desambiguación (y II) 157

Capítulo 1

Introducción

El desarrollo de las computadoras y, su convergencia con las telecomunicaciones ha provocado una revolución en la información, en lo que se ha venido en llamar "sociedad de la información". Como consecuencia de la expansión de las nuevas tecnologías de la información y de las comunicaciones, cada vez hay más gente que diariamente trata con computadoras de manera directa o indirecta. Muchas de las tareas que realizan, implican de un modo o de otro, el uso y tratamiento de la lengua. La redacción y corrección de documentos, la consulta a distancia de fuentes de información, el uso de diccionarios y enciclopedias, la traducción, el envío y recepción de mensajes, son algunas de las actividades en que la lengua representa un papel primordial. Así, llegará a ser normal, no sólo la información presentada con imágenes y sonidos, sino en lenguaje natural, tanto escrito como hablado. Esto es, el comienzo de la edad de la información, donde ésta es vital para el desarrollo económico, social, político de los pueblos, así como para su calidad de vida.

Por otra parte, la cantidad de información a la que una persona puede tener acceso crece exponencialmente, gracias a la redes de computadoras, en especial a Internet, y aunque el tipo de esta información es cada vez más variado, la información textual hoy por hoy es la predominante. Actualmente, sobre el 90% de la información de las corporaciones se encuentra en formato de texto [Oracle, 1997]. Podemos encontrar texto en documentos, manuales, informes, circulares, correos electrónicos, faxes y también en páginas Web. Sólo para este último medio, hay estimaciones [Baeza-Yates y Ribeiro-Neto, 1999] en relación con la cantidad de texto disponible, del orden de un terabyte. La revolución continua y, una de sus consecuencias es el gran volumen de información que se compilará de forma más natural para los usuarios que la información estrictamente estructurada. Esto conlleva a un considerable agravamiento de la sobrecarga de información.

Existen grandes problemas, como el manejo de la ingente cantidad de información, debido a la imposibilidad de identificar, recuperar y seleccionar lo que realmente se necesita. La Ingeniería Lingüística[1] puede ayudar a solventar estos problemas inherentes de recuperación de información

[1] La Ingeniería Lingüística podría definirse, siguiendo un documento de la Comisión Europea, como *la aplicación*

no estructurada y distribuida a través de Internet. El término *Ingeniería Lingüística* abarca un amplio espectro de actividades que suelen englobarse dentro de lo que se ha denominado "las industrias de la lengua". Éstas se centran en "una serie de actividades comerciales en las que el tratamiento del lenguaje por personas o por máquinas o por una combinación de unas y otras, forma una parte fundamental del producto o servicio" [Europea, 1997]. La investigación desarrollada en torno a la Ingeniería Lingüística, tiene como objetivo la consecución de sistemas informáticos que implementen funcionalidades próximas a la comprensión del lenguaje y a la generación del lenguaje humano.

Existe un problema inherente de ambigüedad en estas funcionalidades, lo que ha suscitado un gran interés desde los primeros tiempos del procesamiento del lenguaje natural (PLN) y, se produce cuando se asocian varios sentidos o significados a una palabra (forma léxica). La desambiguación[2] automática es tratada con especial interés dentro del campo de la Ingeniería Lingüística. En un número importante de aplicaciones del PLN, aquellas que son sensibles a la denotación semántica, la introducción de un método de desambiguación automático puede resultar beneficioso.

El objetivo de la desambiguación es identificar el significado correcto de una palabra de los proporcionados en un diccionario, thesaurus o similar. La ambigüedad se produce al existir palabras con varios significados, y puede ser descrita como sigue: una persona comprende una frase con una palabra ambigua, la comprensión se realiza sobre la base de uno de sus posibles significados. Así, el significado apropiado se selecciona sobre un rango de posibilidades, como parte de un proceso de comprensión del lenguaje humano. De esta manera, la desambiguación puede verse como una tarea bien definida, emprendida por un módulo del procesador del lenguaje humano. Este módulo podría ser modelado computacionalmente en un programa de desambiguación

La investigación desarrollada desde los primeros tiempos en desambiguación [Lesk, 1986] hasta la actualidad, ha tratado principalmente la construcción de sistemas automáticos, con distintos enfoques, para la resolución de la ambigüedad, pero sería deseable investigar más en su aplicación a tareas donde la resolución de la ambigüedad léxica pueda ser útil.

Los dos criterios más importantes de mejora que se pretenden en la desambiguación son la efectividad o calidad con que se realiza, y la eficiencia del proceso. No hay una gran convergencia en los métodos de evaluación de la efectividad en la literatura sobre la resolución de la ambigüedad léxica (Word Sense Disambiguation —WSD—). Las medidas más comunes sobre efectividad son las tradicionales en el campo de la Recuperación de Información [Lewis, 1992]. La eficiencia del proceso (a nivel de complejidad de algoritmos y de número de pasadas sobre el texto) es importante, dado que en las aplicaciones se prevé procesar grandes volúmenes de texto en tiempo real.

La desambiguación se puede utilizar para mejorar diferentes tareas de clasificación automática. De hecho, la resolución de la ambigüedad léxica puede considerarse como una "tarea

de los conocimientos sobre la lengua al desarrollo de sistemas informáticos que puedan reconocer, comprender, interpretar y generar lenguaje humano en todas sus formas [OEIL, 1998].

[2]Emplearemos los términos "desambiguación" y "desambiguador" al ser utilizados por otros investigadores.

intermedia" [Wilks, 1998], no un fin en sí misma, pero realmente necesaria para acompañar a muchas y variadas tareas propias del procesamiento del lenguaje natural. La desambiguación es esencial para diversas aplicaciones de comprensión del lenguaje y análisis de documentos, tales como la comprensión de mensajes y comunicación persona-ordenador; y puede introducir mejoras en los siguientes campos de aplicación:

- Recuperación de Información (*Information Retrieval* —IR—),
- Recuperación de Información Multilingüe (*Cross-Language Information Retrieval* —CLIR—),
- Extracción de Información (*Information Extraction* —IE—),
- Categorización de Textos (*Text Categorization* —TC—)
- Traducción automática (*machine translation* —MT—).

Estas tareas padecen los efectos producidos por las palabras con múltiples sentidos. Así los sistemas de recuperación pueden recuperar documentos que contienen las mismas palabras utilizadas en la consulta pero con diferentes sentidos. Por ejemplo, cuando buscamos referencias judiciales, sería conveniente eliminar los documentos recuperados que contienen la palabra "court" relacionada con "royalty". En el proceso de traducción automática se producen situaciones en las que una palabra simple con varios significados puede tener varias traducciones dependiendo del contexto. Por ejemplo, en inglés la palabra "duty" tiene dos significados distintos "tax" y "obligation", su traducción en francés corresponde con "droit" y "devoir" respectivamente.

En esta monografía estudiamos los recursos lingüísticos para la resolución de la ambigüedad, presentando un nuevo modelo basado en la integración de recursos lingüísticos para su utilización y aplicación a la tarea de categorización automática de textos, así como a la recuperación de información.

Introducimos mejoras en los métodos de desambiguación automática de textos, empleando para ello de manera fundamental información procedente de recursos lingüísticos. La hipótesis básica de trabajo es que cuanto más informado esté un sistema mayor efectividad tendrá. Para ello, se propone la utilización de distintos recursos lingüísticos, como corpora de textos, y bases de datos léxicas para realizar el proceso de resolución de la ambigüedad léxica. En concreto, utilizaremos como corpus etiquetado SemCor (Semantic Concordance), debido a su disponibilidad y amplia cobertura. SemCor está compuesto por el Brown Corpus, etiquetado o anotado manualmente con los sentidos de las palabras definidas en WordNet [Miller, 1990, 1995]. Como base datos léxica utilizaremos WordNet. Está organizada por medio de relaciones semánticas, y cuya principal unidad estructural es el "synset" o conjunto de sinónimos, que representa un significado concreto.

En nuestro trabajo, diseñamos métodos WSD para tareas específicas de clasificación automática de documentos y mejorar su efectividad, ya que existe un problema inherente de ambigüedad

de términos. La clasificación de documentos consiste en asignar clases o categorías previamente existentes a documentos [Lewis, 1992]. La categorización de textos ha estado históricamente unida a la recuperación de textos o de información [Salton, 1989]. Esta aplicación es de gran importancia en el ámbito bibliotecario y de documentación, pero en la actualidad se conocen muchas otras aplicaciones de la clasificación de documentos [Lewis, 1992; Hearst, 1994; Wiener et al., 1995], dentro de este tipo de sistemas podemos incluir los sistemas de recuperación de información que seleccionan, en grandes bases de texto, aquellos textos o documentos que son adecuados a una necesidad del usuario. Otro ejemplo lo constituyen los de encaminamiento y filtrado de texto, este tipo de sistemas pueden determinar, a partir del análisis de contenido del texto de un mensaje, cuál es la dirección más adecuada a la que debe enviarse. Pueden incluirse en sistemas de gestión de correo electrónico y artículos de noticias u otros canales continuos de texto con destino a usuarios interesados en ellos.

El orden de exposición que seguimos en el desarrollo de esta memoria es el siguiente:

En el Capítulo 2 se exponen y estudian las distintas tareas del análisis del contenido textual. Después se presentan una serie de elementos y técnicas de indexación que constituyen la base para el desarrollo del sistema de desambiguación, tratando con especial detalle el modelo del espacio vectorial. Finalmente, se exponen otros elementos adicionales relacionados con la indexación automática de los documentos y consultas.

En el Capítulo 3 se exponen y clasifican los distintos recursos lingüísticos, centrándonos fundamentalmente en los corpora de textos y las bases de datos léxicas, particularmente en el corpus de texto SemCor y la base de datos léxica WordNet, puesto que son los que se emplean en el nuevo enfoque propuesto de resolución de la ambigüedad. Asimismo, se exponen otros recursos que son empleados en la tareas de clasificación de documentos que abordamos en el Capítulo 5.

En el Capítulo 4 se analiza el problema de la ambigüedad léxica y su resolución, concebida esta última como un proceso cuyo aspecto fundamental es el análisis de los documentos textuales. Asimismo, se proporciona una perspectiva general del trabajo realizado en desambiguación y se hace una revisión y clasificación de los métodos empleados en los distintos sistemas de desambiguación. Por otra parte, se discuten todos los aspectos relacionados con la efectividad del proceso de desambiguación. Se presentan un enfoque desambiguación basado en corpus y otro basado en base de datos léxica, que materializan en un sistema el nuevo modelo de desambiguación basado en la integración de recursos lingüísticos. A continuación se detalla el planteamiento y la organización del sistema, fundamentado en el modelo del espacio vectorial. Finalmente, se describe el entorno experimental mostrando los recursos empleados e interpretando los resultados de la evaluación directa de la desambiguación centrada en la efectividad.

En el Capítulo 5 se estudia la aplicación de la desambiguación a tareas de clasificación de documentos, en concreto a la recuperación de información y categorización de textos. Para el caso de la recuperación de información se trata la técnica de realimentación como mecanismo de adquisición de información contextual. Se utiliza esta técnica para facilitar la desambiguación en el proceso de expansión de consultas con términos de WordNet y mejorar la efectividad de

la recuperación. Finalmente se describe el entorno de evaluación y se interpretan los resultados obtenidos. En nuestro estudio concluimos que la desambiguación proporciona incrementos importantes en la efectividad de los sistemas de clasificación automática de documentos. Después, en el caso de la categorización de textos se presenta un enfoque basado en la integración de recursos lingüísticos a través de la resolución de la ambigüedad, es decir, aplicamos el enfoque de desambiguación propuesto en el capítulo anterior. Una conclusión global a la que se llega en esta tarea, es que la desambiguación es necesaria para realizar el proceso de integración de recursos lingüísticos de forma automática, y en general necesaria para la tarea propiamente dicha. A continuación, se exponen los experimentos que se han realizado en categorización, centrados en la efectividad, de manera que nos permita medir la incidencia de la desambiguación.

En el Capítulo 6 resumimos las principales aportaciones realizadas y enumeramos las posibles líneas de desarrollo en un futuro inmediato y a más largo plazo. Asimismo, se incluyen cuatro apéndices, tres de ellos con detalles de los recursos lingüísticos utilizados, y con las categorías empleadas tanto en la categorización como en la desambiguación. Otro, con detalles adicionales de los experimentos realizados en desambiguación. Finalmente, se incluye la bibliografía utilizada para el desarrollo de esta memoria.

Capítulo 2

Análisis del contenido

2.1 Introducción

El *Análisis del Contenido Automático* es un conjunto de técnicas para analizar el contenido de los objetos de información y facilitar su posterior acceso [Wilensky, 1999]. Dentro de este proceso genérico se pueden distinguir una serie de tareas específicas, orientadas al procesamiento basado en el contenido de los textos o documentos. Podemos dividir las tareas de procesamiento de texto basadas en el contenido en dos grupos amplios. La *clasificación de texto* involucra la asignación de documentos, o partes de documentos a uno o más grupos de un conjunto de ellos dado. La *comprensión de texto* hace referencia a tareas que involucran la utilización de una mayor cantidad de conocimiento por parte de los sistemas que las implementan (por ejemplo, la generación de resúmenes de texto) [Lewis, 1992; Cowie y Lehnert, 1996].

En este capítulo, describimos las principales operaciones o tareas del análisis del contenido, centrándonos en la clasificación de documentos. No existe una clara frontera definida entre algunas de las operaciones, ya que algunas se solapan en cierta medida, así, los sistemas prácticos pueden utilizar componentes que implementen y combinen varias de ellas. Por otra parte, resulta de especial interés su presentación como representantes de la multiplicidad de las aplicaciones concretas de este tipo de sistemas, y por la innegable utilización de esta terminología en la bibliografía [Salton, 1989; Lewis, 1992; Belkin y Croft, 1992]. En los siguientes puntos se aborda brevemente la desambiguación del significado de las palabras, la traducción automática y la generación automática de resúmenes. Por otra lado, se estudian las técnicas de indexación automática, presentando el modelo del espacio vectorial como la base para la implementación de un gran número de tareas de clasificación de documentos. Asimismo, se trata un conjunto de técnicas en el ámbito del modelo presentado, para realizar el análisis automático del contenido y obtener una representación, como son el cálculo de pesos de términos, las listas de parada, la extracción de raíces y la realimentación. Por último, se incluye un resumen y conclusiones.

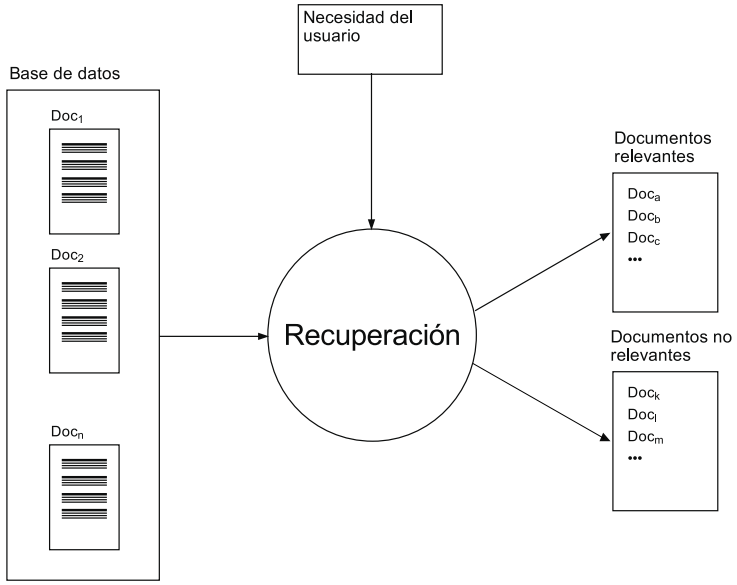

Figura 2.1: Recuperación de textos

2.2 Tareas

2.2.1 Clasificación de documentos

La *clasificación* se refiere al proceso de agrupación de entidades. La clasificación de documentos o de textos es una expresión adecuada para designar en conjunto, tareas de procesamiento de texto que suelen considerarse distintas, pero que todas ellas involucran la agrupación de entidades [Lewis, 1992]. Vamos a describir seis tareas de clasificación de documentos en esta sección, centrándonos principalmente en dos tareas: recuperación y categorización de texto, al ser éstas objeto de aplicación de la resolución de la ambigüedad léxica.

Recuperación de textos

La recuperación de texto es la selección del subconjunto de documentos adecuados a las necesidades de un usuario entre un conjunto más amplio existente en una base de datos documental [Salton y McGill, 1983]. La necesidad del usuario suele encontrarse representada mediante una consulta formada por los términos o palabras que la caracterizan. Típicamente, el sistema selecciona aquellos textos que contienen en mayor medida los términos de la consulta.

El proceso de recuperación de textos puede concebirse de la forma representada en la figura 2.1. El sistema tiene como entradas fundamentales los documentos existentes en la base de datos y la consulta del usuario. El resultado del proceso es la clasificación por parte del sistema de los documentos existentes en dos grandes grupos: los relevantes e irrelevantes, o lo que es lo mismo, aquéllos que son seleccionados por el sistema como adecuados a la necesidad del usuario,

y aquéllos que no. Opcionalmente, el sistema puede asignar un valor numérico de la relevancia (utilidad de la información para el usuario de acuerdo con su consulta) de los documentos (no sólo booleano).

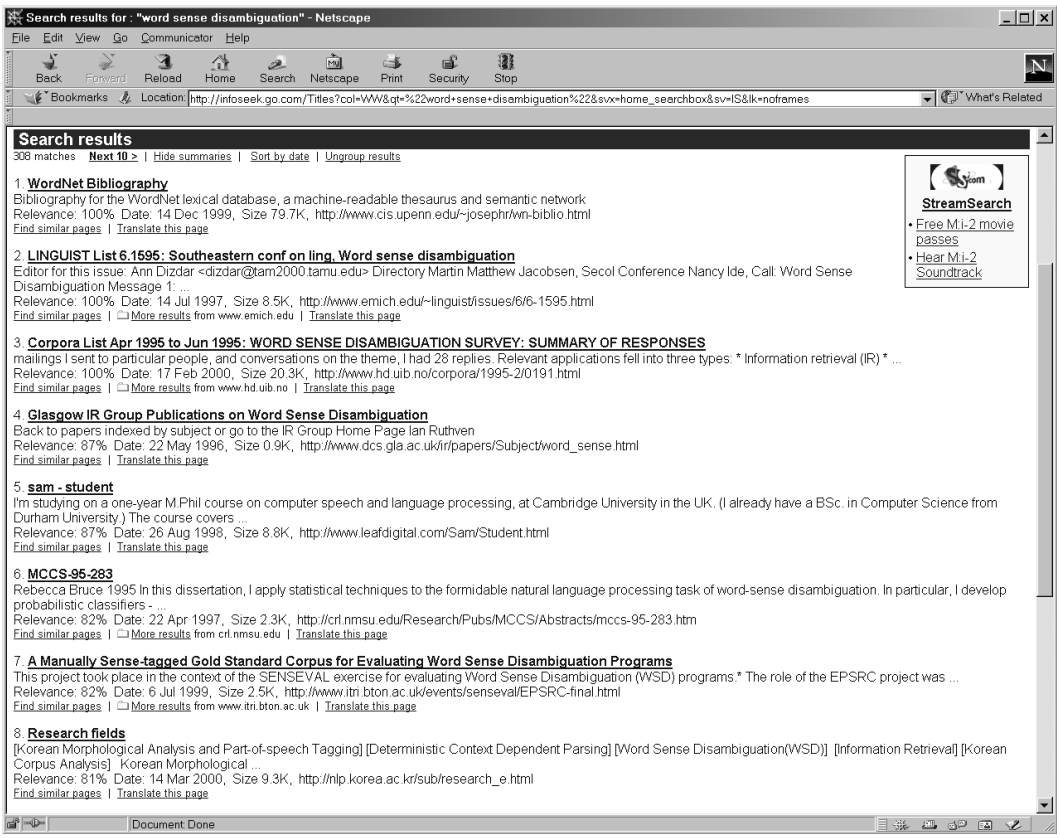

Figura 2.2: Interfaz del sistema infoseek

En la figura 2.2 aparece la interfaz del sistema Infoseek, buscador de información comercial muy extendido en Internet, que implementa técnicas de recuperación de información. El sistema busca los documentos o URL relacionados con la consulta, presentando en el panel de resultados los títulos de los documentos ordenados por la relevancia con dicha consulta. El usuario puede acceder a cualquier documento.

En el ejemplo, se muestra la consulta "Word Sense Disambiguation", como se puede apreciar la interfaz del sistema presenta una lista de los documentos disponibles ordenados por un factor de relevancia.

Categorización de textos

La categorización de textos es la asignación de una o más categorías preexistentes a cada documento [Lewis, 1992]. Un sistema de categorización realiza una labor análoga a un indexador

humano que asigna una serie de descriptores a un documento. Como ejemplo, un sistema de categorización se puede utilizar en un motor de búsqueda de Internet (por ejemplo InfoSeek) para asignar a textos electrónicos disponibles en la red, una o más categorías de entre un conjunto de ellas ya existentes, como cultura, ocio, deportes, universidad, ciencia,... Así, a un informe técnico de un departamento de informática de una universidad, se le podrían asignar las categorías (o descriptores) [universidad, informática, programación] [Vaquero y Buenaga, 1996].

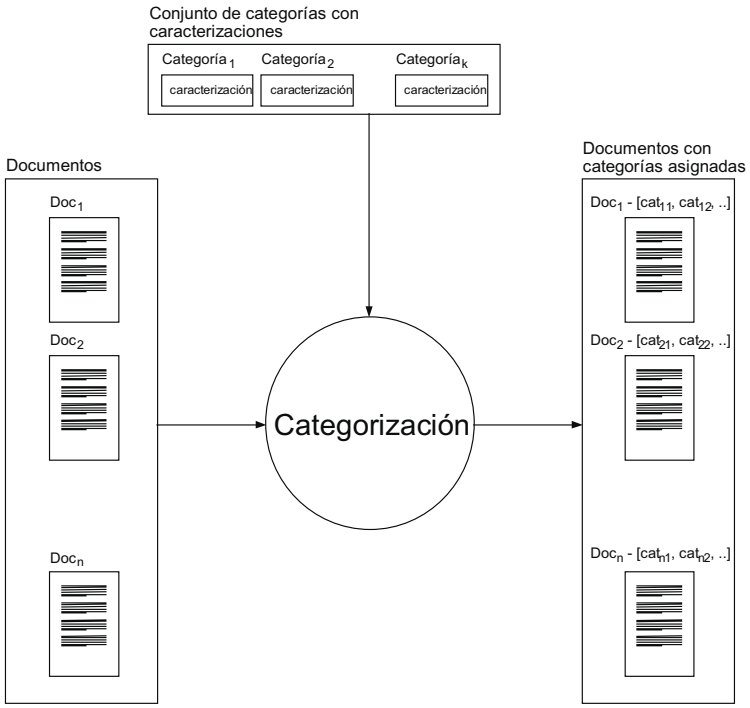

Figura 2.3: Categorización de textos

El proceso de categorización puede concebirse de forma gráfica como se representa en la figura 2.3. Como entradas del sistema se dispone de los documentos y el conjunto de categorías. Como salida se obtienen los documentos con las categorías asignadas. Para cada categoría se dispone de una cierta representación, que puede ser de diferentes tipos. En una aproximación sencilla, se puede disponer de un amplio conjunto de palabras asociadas a cada categoría, en caso de aparecer las palabras de una categoría en un documento, le es asignada.

Encaminamiento de textos

Dado un conjunto de documentos y un conjunto de direcciones a los que deben ser remitidos, el encaminamiento de textos se centra en la asignación de las direcciones adecuadas a cada uno de los documentos [Sheldon, 1995]. Como ejemplo, un sistema de encaminamiento de textos puede ser utilizado para distribuir un conjunto de escritos entre los responsables de los departamentos

de una empresa de forma adecuada. Asimismo, dado un conjunto de documentos con información comercial de tipo publicitario de contenidos diversos (por ejemplo: turismo, electrónica, finanzas, etc.) pueden ser enviados a aquellos usuarios a los que resulten de mayor interés.

Un sistema típico de encaminamiento de textos tiene como entradas fundamentales los documentos a encaminar y las direcciones de destino. Se dispone de una caracterización de los intereses de cada una de las direcciones. Como salida se obtienen los documentos, con la(s) dirección(es) asignadas a cada uno de ellos. La caracterización de los intereses de cada dirección puede ser diversa. Por ejemplo, puede disponerse de un conjunto de términos asociados a cada dirección, que deben aparecer en un documento para que le sea asignada.

Filtrado de textos

El filtrado de textos se centra en la clasificación de documentos llegados por un flujo de información a un destino concreto [Belkin y Croft, 1992]. Como ejemplo, un sistema de recepción de correo electrónico puede incluir un subsistema de filtrado de textos. El sistema filtraría los mensajes llegados al usuario, de forma tal que los mensajes más importantes se encontrarían al comienzo de la lista de mensajes recibidos, mientras que los claramente intranscendentes (por ejemplo: publicidad, listas de distribución masivas), se colocarían al final de los mensajes recibidos, e incluso, opcionalmente, se borrarían de forma automática.

Un sistema de filtrado tiene como entradas fundamentales los textos llegados y una representación del contenido de cada clase que se utilizará para el filtrado. El resultado está formado por los documentos asignados a cada clase. La representación de cada clase utilizada en el filtrado puede ser diversa. Por ejemplo, se pueden incluir una serie de términos para caracterizar la información de especial importancia. En caso de aparecer alguno de ellos en el texto, el sistema lo considerará como tal. De igual forma, pueden tratarse los documentos irrelevantes. El conjunto de clases considerado en un sistema de filtrado puede ser binario (transcendente/intranscendente) o estar formado por un número de elementos mayor, como por ejemplo, los correspondientes a áreas temáticas.

Agrupamiento de textos

El agrupamiento (clustering) de documentos se centra en, a partir de los textos existentes en una base de datos documental, construir automáticamente conjuntos de documentos o textos (y conjuntos de conjuntos) con contenidos semejantes [Salton, 1989; Rasmussen, 1992]. Los grupos construidos pueden facilitar al usuario el acceso a la información. Como ejemplo, un sistema de agrupamiento de texto que trabaja sobre la documentación textual del manual en formato electrónico del sistema operativo UNIX (cada orden dispone de un texto asociado que la describe), puede generar un agrupamiento de las órdenes de acuerdo con su similitud [Maarek y Berry, 1991].

Segmentación de textos

La operación de segmentación se centra en, dado un texto, dividirlo en segmentos, o fragmentos, que tengan una cierta unidad de contenido [Salton y Allan, 1993].

La segmentación resulta de interés para el procesamiento de textos de gran longitud. Como ejemplo, resulta de interés la inclusión de un subsistema de segmentación de textos en un sistema de gestión de una base de datos documental, que dispone del texto de artículos científicos con un número importante de páginas cada uno. Cuando un usuario selecciona un determinado documento, en lugar de serle presentadas las decenas de páginas del artículo de forma lineal, se le presentan los comienzos de los fragmentos en que ha sido segmentado. De esta forma, el usuario puede disponer de una visión global de las partes del contenido del texto. Para la segmentación de un texto se pueden utilizar diferentes tipos de información.

2.2.2 Resolución de la ambigüedad léxica

El proceso de resolución de la ambigüedad léxica consiste en la identificación del significado correcto de un término en un contexto dado, de entre los proporcionados por un diccionario o thesaurus, ya que una misma palabra puede tener diferentes acepciones dependiendo del contexto de que se trate. La distinción de significados más o menos finos depende del diccionario.

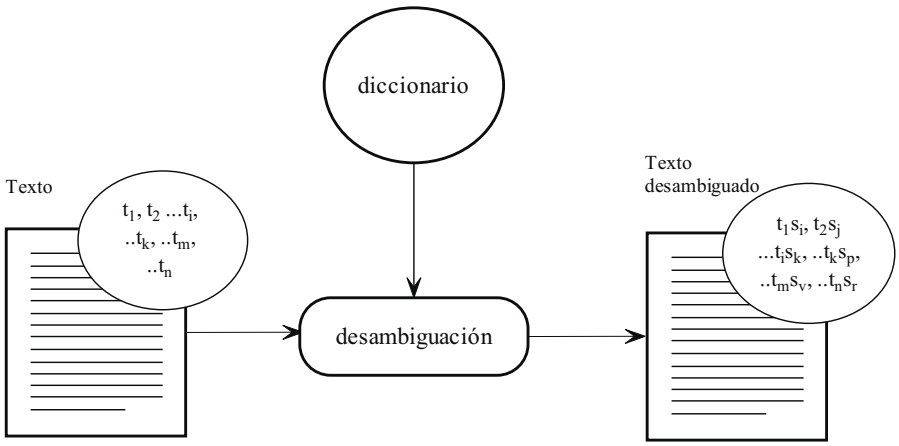

Figura 2.4: Resolución de la ambigüedad

La ambigüedad léxica es quizás un problema importante del PLN. Algunas tareas necesitan determinar el significado correcto para operar eficientemente. La resolución automática de la ambigüedad debería ser una tarea bien definida realizada por un determinado módulo. Este módulo podría ser modelado e implementado en un programa.

Este punto de vista es representado por Cottrell [1989] como sigue:

"Lexical ambiguity is perhaps the most important problem facing a NLU system. Given that the goal of NLU is understanding, correctly determining the meanings of the words used is

fundamental... The task taken here is that it is important to understand how people resolve the ambiguity problem, since whatever their approach, it appears to work rather well..."

De esta manera la resolución de la ambigüedad de las palabras podríamos integrarla como una tarea más en el proceso de análisis del contenido, aunque podríamos considerarla como una tarea intermedia, muy necesaria para algunas de las tareas mencionadas. En la figura 2.4 se describe de forma gráfica el proceso de desambiguación o de resolución de la ambigüedad.

2.2.3 Traducción automática

La traducción automática (*Machine Translation*) es la tarea de traducir de forma automática textos o documentos de un lenguaje natural, que podríamos denominar origen, a otro lenguaje natural destino. Ésta es una tarea difícil debido a la inherente complejidad del lenguaje humano.

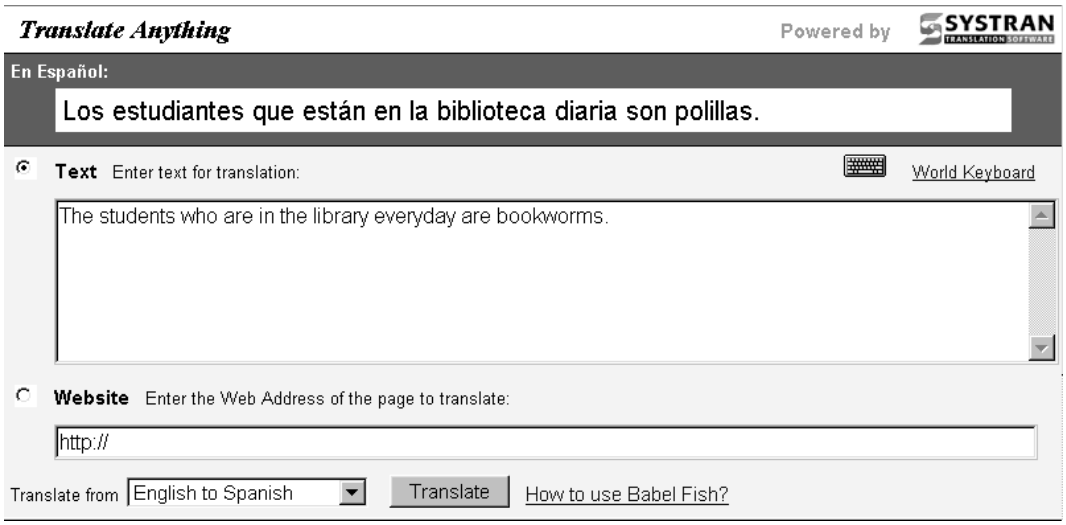

Figura 2.5: Sistema SYSTRAN

El objetivo de un sistema MT es traducir automáticamente de una lengua, hablada o escrita, a otra. La traducción automática resulta especialmente útil para abordar grandes cantidades de texto, pero se precisa actualmente una revisión a fondo para aclarar los significados y los contextos. Los sistemas de traducción automática pueden contribuir de manera significativa en dominios concretos, en los que se emplea un vocabulario limitado, y quizá, en los que se utiliza un serie reducida de estructuras sintácticas. En concreto, proporciona medios útiles de aproximación, que dan al usuario el sentido del documento original, pero de forma no adecuada a su publicación. En definitiva, actualmente la traducción automática sirve de instrumento de ayuda al traductor pero no sustituye a éste.

Un ejemplo de un sistema comercial de traducción automática muy extendido en Internet,

como complemento a muchos sistemas de búsqueda de información, es SYSTRAN[1]. Como se puede observar en la figura 2.5, se propone la traducción de la frase de inglés a español, *"The students who are in the library everyday are bookworms"*. En ella aparece un término ambiguo como bookworm, con dos acepciones bien diferenciadas en español: *polilla* y *ratón de biblioteca*. El sistema no traduce correctamente dicho término en ese particular contexto. Así, si disponemos de un subsistema que resuelva la ambigüedad léxica, mejorará la eficacia del sistema de traducción automática.

Asimismo, resulta de especial interés la traducción automática para la recuperación de información multilingüe (CLIR —Cross-Language Information Retrieval[2]—). Se puede encontrar en [Gachot et al., 1998] una descripción de la implementación del SYSTRAN NLP Browser, único sistema de recuperación de información multilingüe basado en la tecnología de traducción automática.

2.2.4 Generación automática de resúmenes

La generación automática de resúmenes de texto es el proceso automático de obtención de información relevante en función de la tarea a la que está destinada el resumen y de las necesidades (o preferencias) de información de un usuario [Inderjeet y Maybury, 1999]. Se pueden clasificar los resúmenes construidos automáticamente atendiendo a los criterios de propósito (hace referencia al uso potencial del resumen, distinguiendo entre los informativos e indicativos), enfoque (se refiere al ámbito del sumario y puede ser genérico o adaptado al usuario) y alcance (indica si el resumen se realiza sobre un solo documento o un conjunto de ellos) [Thérèse, 1997].

Un sistema de recuperación de información puede incluir un subsistema generador de resúmenes para ayudar al usuario en la selección de un documento respecto a sus necesidades de información, expresadas en forma de consulta en un sistema de recuperación de información.

La salida del sistema de la figura 2.2 tratada anteriormente, al igual que la mayoría de los sistemas de recuperación de información, ampliamente utilizados en Internet, presenta una lista de documentos ordenada por el factor de relevancia y, para cada uno, su título y las primeras palabras del mismo, además de información adicional (tamaño, fecha de creación y lugar donde se encuentra). A partir de este resultado el usuario debería poder decidir cuáles de esos documentos responden a sus necesidades de información plasmadas en la consulta. Aunque en algunos casos la evidencia es importante, en otros no puede tomarse la decisión sin examinar el documento completo. La consecuencia es que se incrementa de manera notable el coste de la operación de recuperación.

[1] http://babelfish.altavista.com/translate o http://www.systransoft.com/

[2] Se refiere a una tarea especial de recuperación de texto, donde la consulta se formula en una lengua (fuente) y los documentos se encuentran en diferentes lenguas al actual, así en el proceso de recuperación no se tienen en cuenta, ni la lengua en que se encuentran los documentos, ni en la que han sido formuladas las consultas. Por tanto, la idea es recuperar todos los documentos relevantes a una consulta, sin considerar idiomas particulares de cada documento y consulta.

2.3 Técnicas de indexación automática

La indexación se suele denominar al proceso de obtención de una representación interna de los documentos [Salton y McGill, 1983; Croft, 1993]. La representación es un problema importante en los sistemas de clasificación de documentos. Para la implementación del proceso de indexación es necesario definir, primero, una forma de representación (o lenguaje de representación) de la información contenida en los documentos y después, desarrollar el analizador que permita obtener las representaciones concretas de los documentos a partir del análisis de su contenido. La representación más idónea es aquélla que permite realizar el análisis de los documentos automáticamente. Este tipo de representación se basa en términos o palabras clave. Una vez realizados los procesos de indexación y análisis, se debe realizar la comparación para obtener el grado de similitud a partir de sus representaciones, dependiendo de la tarea de clasificación de documentos.

A continuación se introducen los principales elementos que se pueden utilizar para la definición de métodos de indexación. En concreto el modelo del espacio vectorial que permite realizar el proceso de indexación empleando los pesos de términos, las listas de parada y la extracción de raíces. Históricamente, estos elementos fueron definidos para la implementación de los sistemas de recuperación de texto.

2.3.1 El modelo del espacio vectorial

El modelo del espacio vectorial proporciona la base para la implementación de un gran número de tareas de clasificación de documentos tales como la recuperación de información, el agrupamiento de documentos, la categorización de textos y el encaminamiento de texto [Lewis, 1992].

El modelo del espacio vectorial ha constituido la base de gran parte de los experimentos y sistemas desarrollados en el campo [Salton y McGill, 1983; Salton, 1991a; Harman, 1992b; Lewis, 1992; Buenaga et al., 1997]. En concreto, en recuperación de información, el modelo de espacio vectorial [Salton y McGill, 1983; Salton, 1989] se encuentra entre los métodos de representación más utilizados. Su efectividad, respecto a otros enfoques como el modelo probabilístico [Rijsbergen, 1979], y otros modelos basados en distintas técnicas como redes bayesianas [Croft, 1991], redes neuronales [Scholtes, 1993], o algoritmos genéticos [Holland, 1992] ha quedado demostrada en trabajos como el de Salton [1989].

Además, el modelo del espacio vectorial define un método de cálculo de similitud y, permite obtener fácilmente una representación interna a partir de un análisis automático de sus contenidos. Esto unido a su extendida utilización ha constituido la razón por la que hemos tomado el modelo para desarrollar nuestro estudio.

Para la tarea de recuperación de textos (información) se pueden distinguir un conjunto de subtareas:

Indexación adecuación y representación de los textos o documentos para permitir un eficiente almacenamiento y una comparación con las consultas. Para tener una estructura interpre-

table por el sistema de recuperación, los documentos deben ser convertidos a expresiones en alguna representación de texto.

Formulación de la consulta o procesamiento (posiblemente) dentro de un lenguaje de consulta intermedio más o menos sofisticado. El usuario debe formular sus necesidades en forma de consulta interpretable por el software de recuperación. La consulta puede ser introducida de una manera similar a la utilizada por el sistema de recuperación, tales como expresiones booleanas. En otros casos, la consulta se puede introducir en lenguaje natural o mediante un documento de ejemplo, para lo cual el sistema debe seleccionar palabras importantes de la entrada del usuario.

Comparación de la consulta con los documentos para obtener factores de relevancia. El sistema debe comparar implícita o explícitamente la consulta del usuario con los documentos almacenados, y hacer una clasificación sobre qué documentos se recuperan y cuáles no.

Realimentación *(feedback*[3]*)* de la consulta para mejorar la efectividad de la recuperación. Raramente una recuperación inicial de documentos resulta adecuada a las necesidades del usuario, para conseguir resultados más aceptables son necesarias varias iteraciones modificando o refinando la consulta. Esta modificación la puede realizar explícitamente el usuario, o también éste puede señalar algunos documentos como apropiados y el sistema implícitamente actualizaría la consulta (consulta más refinada).

A continuación, se describe el proceso de representación y cálculo de la similitud empleando el modelo del espacio vectorial para un sistema de recuperación de información.

En el modelo de espacio vectorial tanto los documentos como las consultas se representan mediante un vector de pesos relativos al conjunto de términos escogidos para su representación. De esta forma, el número de términos[4] seleccionados determina la dimensión del espacio vectorial. Así, si suponemos un conjunto de m términos $term_i$, el documento $\vec{d}_j$ se representa mediante el vector cuyas componentes son los pesos asociados wd_{ji}:

$$\vec{d}_j = \langle wd_{j1}, wd_{j2}, ..., wd_{jm} \rangle \tag{2.1}$$

La consulta $\vec{q}_k$ se representa, de manera análoga, mediante el vector de pesos wq_{ki} relativos al conjunto de términos:

$$\vec{q}_k = \langle wq_{k1}, wq_{k2}, ..., wq_{km} \rangle \tag{2.2}$$

[3]A este proceso nos referiremos en el Capítulo 5 como *relevance feedback*.

[4]La definición más común de un término (en inglés, además de otras muchas lenguas románicas), y la que utilizamos en esta memoria, es que un término es una secuencia de caracteres alfanuméricos delimitados por espacios en blanco (espacios, tabuladores o nuevas líneas) signos de puntuación (tales como una coma). Además, las letras mayúsculas se ignoran, ya que todas ellas en un documento se cambian a minúsculas.

A partir de este método de representación se calcula la similitud entre la representación de un documento $\vec{d_i}$ y la de una consulta $\vec{q_k}$: basta medir el coseno del ángulo que forman ambos vectores en el espacio m-dimensional. La expresión asociada es la siguiente:

$$sim(\vec{d_j}, \vec{q_k}) = \frac{\sum_{i=1}^{m} wd_{ji} \cdot wq_{ki}}{\sqrt{\sum_{i=1}^{m} wd_{ji}^2 \cdot \sum_{i=1}^{m} wq_{ki}^2}} \tag{2.3}$$

Mediante esta expresión, se asigna un valor de similitud a cada documento respecto a la consulta del usuario. Este valor permitirá al sistema de recuperación, ordenar el resultado de la búsqueda en función del grado de relevancia[5], como se observa en el ejemplo de la figura 2.2, a fin de ayudar al usuario a localizar la información que necesita [Harman, 1992b].

Se pueden utilizar un conjunto de técnicas en el ámbito del modelo de espacio vectorial [Salton, 1989], que permiten el análisis automático del contenido para la obtención de la representación de documentos y consultas: cálculo del peso de los términos, listas de parada y extracción de raíces.

2.3.2 Pesos de términos

Se han desarrollado diversas técnicas para el cálculo del peso de los términos que deben representar internamente a los documentos. La mayor parte estos métodos se basan en el concepto de *poder de resolución* de un término[6], entendido como medida de su adecuación para ser término de indexación [Salton y McGill, 1983; Rijsbergen, 1979] y, para reducir la dimensión de los vectores de documentos. Luhn [1958] establece una relación entre el grado de discriminación o poder de resolución de un término y su frecuencia de aparición en el documento. Así las palabras con mayor poder de resolución tienen una frecuencia de aparición media. La justificación para la eliminación de términos infrecuentes se basa en una observación, realizada por Zipf [1949] y conocida como "Ley de de Zipf[7]", sobre la frecuencia de aparición de las palabras en un corpus de textos, establece que, ordenadas las palabras de un texto (o conjunto de textos) por su frecuencia de uso, el producto de su frecuencia de uso por su posición en el ordenamiento es constante. Esta relación se muestra gráficamente en la figura 2.6.

Una forma de conseguir una asignación de pesos próxima a la definición de poder de resolución es la que se propone en [Salton, 1989]. Se basa en la asignación de pesos a los términos que aparecen en los documentos mediante las expresiones:

$$wd_{ji} = tf_{ji} \cdot idf_i \tag{2.4}$$

$$idf_i = \log_2 \left(\frac{n}{df_i} \right) \tag{2.5}$$

[5]Como se puede extraer, el valor de la similitud entre un documento y una consulta es la relevancia calculada por el modelo.

[6]Proporciona una base para los métodos de indexación basados en frecuencias de aparición de términos.

[7]Actualmente no es una ley sino meramente un fenómeno empírico.

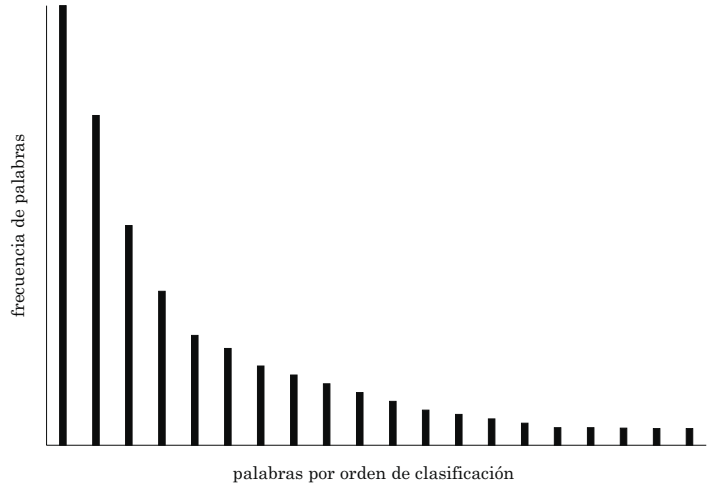

Figura 2.6: Histograma de la frecuencia de las palabras mostrando la ley de Zipf

Donde wd_{ij} es el valor del peso que se le asigna al término i en el documento j, tf_{ji} la frecuencia de aparición en el documento j del término i e idf_i es la inversa de la frecuencia de documento para el término i. En la expresión que calcula este último valor, n indica el número de documentos de la colección y df_i, la frecuencia de documento del término i, entendida como el número de documentos en la colección en que aparece dicho término.

2.3.3 Listas de parada

Como se ha comentado en la sección anterior, hay un conjunto de términos de uso muy frecuente que carecen de poder de resolución y, por tanto, no son buenos candidatos para formar parte de un índice de términos (términos de indexación). Para la eliminación de aquellos términos que no aportan poder de resolución, es decir, que no son útiles para el proceso de indexación, se vienen utilizando desde los comienzos de IR, las denominadas listas de parada (*stoplists*) [Fox, 1992] para aumentar la efectividad del proceso de recuperación de información. Ejemplos de estas palabras en inglés, con poca capacidad de discriminación y poco representativas como términos de indexación, son: "the", "of", "and", "to" ...

Estas palabras vacías son muy frecuentes en los documentos, como lo demuestra el hecho de que las diez palabras de más frecuente uso en inglés, pueden llegar a ser entre el 20% y el 30% por ciento de los términos de un documento [Francis y Kucera, 1982].

Estas listas se obtienen en estudios orientados específicamente a ello, a partir de un corpus de textos lo suficientemente representativo del idioma considerado. Por ejemplo, puede encontrarse en [Rijsbergen, 1979] una lista de parada de 250 términos y en [Fox, 1992] una de 425 obtenida a partir del Brown Corpus [Edwards y Lampert, 1992].

Para evaluar la aplicación de nuestro desambiguador a la recuperación de información utiliza-

a	all	and	are	became
a's	allow	another	aren't	because
able	allows	any	around	become
about	almost	anybody	as	becomes
above	alone	anyhow	aside	becoming
according	along	anyone	ask	been
accordingly	already	anything	asking	before
across	also	anyway	associated	beforehand
actually	although	anyways	at	behind
after	always	anywhere	available	being
afterwards	am	apart	away	believe
again	among	appear	awfully	below
against	amongst	appreciate	b	beside
ain't	an	appropriate	be	besides

Tabla 2.1: Primeros términos de la lista de parada utilizada en SMART

mos la misma lista que emplea el sistema SMART[8], formada por 571 términos, y cuyas primeras palabras se muestran en la tabla 2.1. En cuanto a la implementación, se ha optado por incluir la eliminación de las palabras de parada como parte del analizador léxico al ser más eficiente. Por tanto, se ha utilizado un algoritmo generador de analizadores léxicos para la indexación automática [Fox, 1992].

2.3.4 Extracción de raíces

La extracción de raíces (*stemming*) es una técnica utilizada para aumentar el rendimiento de los sistemas de IR, ya que aumenta la efectividad del proceso de recuperación y reduce el tamaño de los archivos de indexación[9]. El objetivo de los algoritmos de extracción de raíces, o de eliminación de sufijos, es obtener un único término a partir de diferentes palabras con un mismo significado que difieren esencialmente en su morfología [Frakes, 1992; Krovetz, 1993].

Como ejemplo, podemos considerar la obtención del término INFORM a partir de "information", "inform", "informer" e "informed". El resultado del algoritmo debe ser una misma forma canónica para las diferentes variantes morfológicas de una palabra, que no tiene por qué ser, necesariamente, la raíz lingüística. En la bibliografía se pueden encontrar diferentes tipos de algoritmos de stemming[10] [Frakes, 1992]. Se pueden introducir dos tipos de errores en este tipo de algoritmos:

[8] Sistema de recuperación de información creado en la Universidad de Cornell por Salton [Salton, 1991a].

[9] El factor de reducción es del orden del 50% [Frakes, 1992]

[10] Uno de los más conocidos es el de Porter [1980]. Se caracteriza por su pequeño tamaño, así como por la simplicidad del conjunto de reglas para eliminar los sufijos. Este simple algoritmo sustituye satisfactoriamente un análisis morfológico para el proceso de recuperación de texto.

- error de *infrarradicación* o *understemming*, que resulta de obtener diferentes formas canónicas o raíces para palabras, que deberían proporcionar una misma por tener un mismo significado; por ejemplo, para "habit" y "habitable" se obtiene HABIT respectivamente.
- error de *sobrerradicación* u *overstemming* se produce cuando se obtiene la misma forma canónica para palabras que deberían tenerlas distintas, por diferir no en variaciones morfológicas, sino en su significado; por ejemplo, para "capital", "capitulate" y "capitol" obtener CAPIT.

Para evitar errores en el proceso de recuperación interesan algoritmos que minimicen la sobrerradicación e infrarradicación. Algoritmos menos sofisticados que tiendan al overstemming y eviten el understemming (se disminuye el número de índices) pueden adecuarse a sistemas en que aspectos de eficiencia en tiempo y espacio representan un papel prioritario, pero conllevan una disminución en la efectividad del proceso [Frakes, 1992].

En la aplicación de nuestro sistema desambiguador hemos utilizado el *stemmer* que incluye SMART.

2.3.5 Realimentación

También se puede utilizar en los sistemas de recuperación de texto el concepto de realimentación [Harman, 1992b,a] para mejorar la eficacia de los sistemas de IR. Esto implica permitir al usuario formular su consulta mediante la selección de una serie de documentos de la base de datos que él especifica al sistema que son adecuados a sus necesidades. Este paradigma de especificación de la consulta del usuario podría resumirse en "encuentra más documentos semejantes a este (o estos)". Para la implementación de la realimentación, se construye el vector de la consulta como la suma de los vectores de los documentos que el usuario ha especificado que son relevantes, esto es:

$$\vec{q}_k = \sum_{i=1}^{n} \vec{q}_i \cdot \alpha(k,i) \tag{2.6}$$

En donde $\alpha(k,i)$ proporciona un valor 1 ó 0 dependiendo de que el documento d_i haya sido especificado como relevante por el usuario a la consulta q_k, o no, respectivamente.

2.4 Resumen y conclusiones

En este capítulo hemos hecho una presentación breve de las tareas de análisis del contenido textual, describiendo los aspectos que tienen una mayor relevancia para este trabajo.

Hemos seleccionado para la desambiguación unos elementos utilizados en indexación dentro del proceso de recuperación de información. Éstos son, en concreto: el modelo del espacio vectorial, la utilización de pesos de términos basados en frecuencias de aparición, uso de listas de parada y algoritmos de extracción de raíces. La utilización de estos elementos introducidos

en este capítulo proporcionan la base para la implementación del proceso de desambiguación y su aplicación a dos tareas de análisis de contenido, la categorización de textos y la recuperación de información, de forma tal que, hoy por hoy, la efectividad conseguida es difícil de superar por otras aproximaciones.

Capítulo 3

Recursos lingüísticos

3.1 Introducción

A comienzos de los años 90 resurge el interés por los métodos estadísticos para el análisis del lenguaje al estilo de los utilizados en los 50. En dicha década, el empirismo dominó un amplio conjunto de campos que fueron desde la psicología (conductivista) a la ingeniería eléctrica (teoría de la información). En ese momento, una práctica común en la lingüística, fue la de clasificar las palabras no sólo, por su significado, sino también por su concurrencia con otras palabras. Firth [1957] resumió en una frase que sintetiza este enfoque:

"You shall know a word by the company it keeps"

Lamentablemente, este interés fue desapareciendo poco a poco al final de esta década y principios de la siguiente [Chomsky, 1957].

Quizás la razón más inmediata de este renacer empirista fue la disponibilidad de cantidades masivas de información en forma de corpora textuales[1] y lexicones. Hoy día, ingentes cantidades de recursos que podemos denominar *recursos lingüísticos* están disponibles gracias a los esfuerzos de distintas organizaciones y asociaciones (Association for Computational Linguistics' Data Collection Initiative —ACL/DCI—, European Corpus Initiative —ECI—, British National Corpus —BNC—, Linguistic Data Consortium —LDC—, Instituto Cervantes, etc.).

La expresión *recursos lingüísticos* se refiere a un conjunto generalmente extenso de datos, así como a descripciones de una lengua en formato electrónico, empleados para mejorar y evaluar los sistemas de procesamiento del lenguaje natural. Ejemplos de recursos lingüísticos[2] son los corpora de textos (escritos y orales), los lexicones (bases de datos léxicas, thesaurus, y diccionarios electrónicos).

[1] Más textos disponibles como la compilación y disponibilidad del Brown Corpus [Francis, 1982].

[2] Aunque este término puede ser extendido a herramientas de software básico para la preparación, colección, gestión, o uso de otros recursos.

La organización de este capítulo es como sigue. En primer lugar, se estudian los corpora de textos, y se establece una tipología dependiente del propósito de los mismos y, una clasificación atendiendo a su contenido. Seguidamente, se estudian las características de un corpus y se describen particularmente dos corpora de textos, SemCor y Reuters ya que son utilizados en nuestros sistemas. En un segundo lugar, se estudian los lexicones, encuandrando dentro de ellos a las bases de datos léxicas, prestando especial atención a WordNet, por ser un recurso lingüístico base para nuestra investigación en WSD. Asimismo, se tratan los diccionarios electrónicos y los thesaurus. Para acabar el capítulo se incluye un resumen y conclusiones.

3.2 Corpora de textos

Un corpus lingüístico es una colección de textos representativos [3] de una lengua, de un dialecto o un subconjunto de un lenguaje, que son utilizandos para el análisis lingüístico [Francis, 1982]. Aijmer y Altenberg [1991] lo describen de forma más simple como:

"... study of language on the basis of text corpora."

Aunque la compilación de los primeros corpora se inicia en los años 60, los corpora lingüísticos han alcanzado su mayor popularidad en estos últimos años, debido principalmente al éxito de los métodos estadísticos, así como al incremento de los sistemas informáticos en cuanto a cálculo y capacidad de almacenamiento. Esto ha favorecido, por un lado, la investigación lingüística en una mera adaptación de sus técnicas, y por otro, la potenciación y productividad del área del procesamiento del lenguaje natural. Todo ello, gracias al empleo satisfactorio de los métodos estadísticos en el PLN, y a la existencia de grandes cantidades de texto en formato electrónico.

Muchas clasificaciones de los corpora son posibles dependiendo del criterio que se considere. Atendiendo al material que incluyen se pueden clasificar en dos grandes grupos: textuales y orales[4], dependiendo si son recopilaciones de texto escrito o de transcripciones de una lengua.

Una tipología básica de los distintos corpora textuales se puede establecer dependiendo del propósito: corpora con fines generales, cuyo objetivo principal es el de constituir una fuente de información textual de una lengua para fines y aplicaciones diversas; y corpora con fines específicos, creados en respuesta a un propósito particular, como el estudio de aspectos concretos de la gramática o del léxico de la lengua, la extracción de datos estadísticos, el estudio del

[3] Leech [1991] entiende por representativo *"... a corpus is representative to the extent that findings based on its contents can be generalized to a larger hypothetical corpus"*. Por ejemplo, se asume frecuentemente que el Brown Corpus es representativo del inglés americano, del inglés escrito o del inglés en general.

[4] Los corpora de textos orales son aquellas recopilaciones de materiales cuyo objetivo principal es caracterizar desde un punto de vista lingüístico la lengua hablada. Por este motivo, consisten en transcripciones ortográficas de grabaciones realizadas en diversas situaciones de uso de la lengua. Las transcripciones pueden estar codificadas de acuerdo con algún formato estándar (por ejemplo el de la TEI, Text Encoding Initiative) y presentar, además, distintos niveles de anotación. Este tipo de corpus es un recurso básico y se compila para el desarrollo de aplicaciones en el ámbito de las tecnologías del habla.

comportamiento lingüístico de una determinada población de hablantes, análisis comparativos de diversas variedades lingüísticas, o el desarrollo y evaluación de sistemas de procesamiento del lenguaje natural. Conviene tener en cuenta que un corpus creado con fines específicos puede muy bien ser reutilizado para tareas distintas de las previstas, siempre y cuando sus características y diseño se adecúen a ellas.

Según el contenido del corpus, se pueden establecer dos grandes categorías: corpora de la lengua general, que reúnen diversos tipos de textos[5]; y corpora de sublenguajes determinados, que limitan su contenido a un tipo concreto de textos.

Finalmente, el corpus puede estar lingüísticamente anotado[6] o no-anotado. Un corpus no-anotado sólo dispone de la colección de textos sin ninguna información adicional. Esta tipología es la más frecuente. Ejemplo de este tipo es el original Brown Corpus [Francis y Kucera, 1982]. Por otro lado, los corpora anotados han supuesto un avance considerable en los últimos años, ya que proporcionan información adicional al texto en forma de marcas o anotaciones incluidas en cada secuencia de caracteres.

Centrándonos en nuestro estudio, podemos precisar que los corpora de textos escritos son colecciones de textos en formato electrónico, que han sido elaborados con fines investigadores y aplicables a variadas tareas del procesamiento del lenguaje natural. Se han empleado en la mejora de los correctores ortográficos y gramaticales, así como en la restauración de acentos, integrándolos dentro de paquetes de procesamiento de texto comerciales. Asimismo, se han propuesto distintos enfoques empleando corpora de textos como recurso lingüístico produciendo mejoras significativas en algunas tareas de análisis del contenido, como resolución de la ambigüedad léxica, recuperación de información, categorización de textos, etc.

3.2.1 Características de un corpus

El tamaño de un corpus se utiliza con frecuencia como medida principal de su calidad. Sin embargo, el tamaño no es una característica definida de un corpus, ni el único parámetro a tener en cuenta. Aunque el tamaño aporta mejoras evidentes en un corpus, la diversidad y el equilibrado de los ejemplos que incluye, o la precisión y variedad en los niveles de anotación son características incluso más significativas para la evaluación de un corpus que el propio tamaño. Leech [1991] señala como prueba de ello que los corpora de tamaño más grandes no son siempre los mejores, pues los pequeños siguen siendo necesarios. En muchos casos, éstos son corpora de carácter específico que se centran en aspectos concretos proporcionando los datos precisos a la tarea de que se trate.

[5] Un ejemplo de este tipo con fines generales es el CREA (Corpus de Referencia del Español [Instituto-Cervantes, 1999]), que está compuesto por textos con distinto carácter (textos literarios, periodísticos, científicos y técnicos, así como transcripciones de grabaciones de la lengua oral y de medios de comunicación en diferentes proporciones).

[6] El término "anotado o etiquetado" fue definido por Souter [1993] como *"... the labelling of a machine-readable text with markers to allow the original text to be accurately re-created on paper in a human-readable form, or to allow new texts created on computer to conform to a common formatting standard, Standard Generalized Mark-up Language (SGML)"*.

Sin embargo, hay que admitir que el tamaño es un parámetro clave en el proceso de construcción de un corpus. El tamaño final del total del corpus radica en la combinación y armonización de un número de aspectos, tales como el propósito, los tipos de textos representados, etc.

El tamaño se relaciona con dos medidas diferentes. Por un lado con el tamaño de la muestra, es decir, las partes de texto de que se compone el corpus. En el tamaño de la muestra ha influido el modelo original establecido por Brown [Sinclair, 1991], y que oscila entre las 2.000 palabras del Brown Corpus y las 45.000 del BNC (British National Corpus) [Burnard, 1995]. Aunque ha sido muy discutido, la conclusión es que el tamaño de la muestra no puede fijarse como un estándar universal. Por otro lado, se relaciona con la dimensión final del corpus. Esta dimensión puede estar determinada por el propósito, ya que las representaciones generales de una lengua necesitan un número considerable de ejemplos de textos variados. Los corpora específicos se centran en un dominio concreto, reduciendo, por tanto la cantidad de ejemplos.

Por otra parte, sería deseable para nuestra tarea, utilizar grandes corpora con ejemplos representativos de la lengua general o de las lenguas de que se trate. Pero si tenemos que elegir entre un corpus grande y un corpus representativo o equilibrado, ¿qué aspecto es más importante para las tareas a realizar?. Mucho se ha discutido sobre este tema [Sinclair, 1991], sin embargo no se ha encontrado una respuesta definitiva. Church y Mercer [1993] establecen que esta cuestión se traduce en el dilema cantidad o calidad: mientras que los laboratorios americanos (por ejemplo IBM, ATT) tienden a favorecer la cantidad, en cambio el BNC, y algunos editores de diccionarios —especialmente en Europa— apuestan por la calidad. Biber [1993] argumenta a favor de la calidad, basándose en que suposiciones inapropiadas o pobres ejemplos pueden producir resultados equívocos. Sin embargo, parece que la conjunción de la calidad y cantidad es más beneficiosa como afirman Church y Mercer después de analizar una decena de corpora: *"more data are better data"*.

Consideramos factores determinantes para lo que podríamos denominar poder de resolución de un corpus en lingüística computacional, la combinación del tamaño de un corpus[7], con su naturaleza dinámica[8], con su amplia disponibilidad y con la capacidad de procesar esto computacionalmente.

3.2.2 Corpora anotados o etiquetados

Los corpora anotados poseen un alto valor lingüístico, no son un mero compendio de palabras, sino que éstas presentan información lingüística en forma de anotaciones. Las anotaciones lingüísticas de un corpus proporcionan información potencial sobre varios niveles del lenguaje

[7]Por tamaño de un corpus consideramos tanto el número de palabras contenidas y medidas en millones, como el número de ejemplos de la lengua que determinan el valor de representatividad y comprensión del corpus.

[8]Se ha planteado una discusión sobre la fuente a utilizar para la adquisición de conocimiento léxico. Por un lado, se han utilizado los lexicones, representados en concreto por los diccionarios, y por otro los corpora. Argumentándose a favor de éstos últimos la naturaleza dinámina de los objetos de la lengua representados, así como destacando su valor como una alternativa para la adquisición de conocimiento léxico [Boguraev y Pustejovsky, 1996].

(*part-of-speech*, sintáctico, etc.). Al principio los esfuerzos realizados en anotación de corpora se centraron en etiquetar las clases de palabras en términos gramaticales con más o menos detalle ("fine" o "coarse grained"). Este interés en etiquetar las clases de palabras sobre la base de descripciones gramaticales, se debe a que estas clasificaciones de elementos lingüísticos son las más elementales y relevantes, así como las más difundidas. Y también, a que son unas de las mejores categorizaciones establecidas en anotación lingüística desde la aparición del primer corpus anotado Brown (el cual contenía etiquetas gramaticales).

Una práctica común en los últimos años ha sido la anotación sólo a un nivel específico, siendo el más frecuente el gramatical, por ejemplo el Brown Corpus está anotado gramaticalmente y no semánticamente. Sin embargo la anotación del mismo corpus se ha realizado a varios niveles en algunos casos. Algunos corpora de este tipo, con anotaciones a varios niveles pueden considerarse "corpora analizados", los cuales son definidos por Sampson [1992]:

"... a sample of natural language annotated not just with grammatical classification of individual words, as in the case of 'tagged corpus', but with codes showing the grammatical structures and perhaps the semantic properties of the material at the levels of phrase, clause, and sentence..."

La proliferación en los últimos años de los corpora anotados hizo necesario un acuerdo en cuanto a los formatos de codificación, al existir una gran variedad de estos. La necesidad de una norma que fuese comúnmente aceptada entre los investigadores, dio origen a TEI (Text Encoding Initiative) que Burnard [1992] describía como un proyecto internacional cuya tarea era:

"... to develop and disseminate a clearly defined format for the interchange of machine-readable texts among researchers, so as to allow easier and more efficient sharing of resources for textual computing and natural language processing [... as well as to make] recommendations about which textual features should be distinguished when encoding texts from scratch, to help ensure that the resulting text can be maximally useful to the research community".

En resumen, TEI contribuye con una guía para la codificación e intercambio de textos en formato electrónico, donde el objetivo primeramente es "cómo" definir un formato para el intercambio de texto entre los investigadores; así como "qué" prácticas específicas se recomiendan en la codificación de nuevos textos. Por tanto, no se trata tanto de una imposición de reglas en el procesamiento del lenguaje para la construcción de corpora, sino de asegurar la compatibilidad del trabajo en este campo con otros recursos lingüísticos, además del establecimiento de recomendaciones que puedan ser seguidas por los investigadores.

Las primeras anotaciones lingüísticas que se realizaron fueron manuales, no así las que se han venido realizando en los últimos años cuya anotación ha sido de forma totalmente automática o semiautomática. Estas anotaciones básicamente muestran la estructura del texto, basándose en la descripción estándar SGML. Utilizan algunas unidades básicas como etiquetas, atributos y referencias a entidades. Las anotaciones automáticas introducen algunos errores en el corpus, mientras que la manual es muy cara en términos de recursos humanos. Una investigación dirigida

a reducir el esfuerzo humano en la anotación de corpora de entrenamiento se presenta en [Engelson y Dagan, 1996]. Consta de algoritmos que seleccionan ejemplos más informativos que podrían ser anotados para luego utilizarse en el entrenamiento. Esta misma idea se presenta en el trabajo de Lehmann et al. [1996], quienes desarrollan una base de datos que contiene ejemplos positivos y negativos de fenómenos lingüísticos diferentes, de manera que pueda construirse un corpus de prueba o entrenamiento sobre algún fenómeno determinado a bajo coste[9].

Como ejemplos de anotaciones que se han venido realizando, podemos considerar los *POS tagging* (etiquetadores), el *parsing* y los anotadores semánticos. De los dos primeros se han construido los denominados *taggers* y *parsers* respectivamente y han permitido etiquetar de manera automática muchos corpora de textos. En cambio las anotaciones de carácter semántico que se incluyen en algún corpus se han realizado generalmente de forma manual y de acuerdo con los significados proporcionados por algún diccionario.

POS tagging consiste en la asignación de etiquetas gramaticales (indicando la parte de la oración: nombre, verbo, determinante, etc.) a cada una las palabras que integran una frase, indicando la función de cada palabra en ese contexto específico. Aunque esto depende de la granularidad del conjunto de etiquetas utilizado —que puede variar de 20 a 500 etiquetas—, esto puede considerarse una tarea fácil, ya que muchas palabras —entre el 80% y 90%— tienen sólo un posible POS, o el contexto en el que aparecen restringe la elección a una sola etiqueta [Padró, 1997]. Pero en el tanto por ciento restante, la resolución de la ambigüedad puede ser dificultosa, ya que muchas veces se requiere de información semántica o incluso del propio sentido común.

Se han construido, como ya se ha comentado, *taggers* con una alta precisión. Los factores que influyen en la precisión de un *tagger* son el conjunto de etiquetas considerado y la manera en que las palabras desconocidas se tratan. Si el *tagger* sigue un enfoque estadístico, el ruido en el entrenamiento y el corpus de prueba representa un importante papel en la efectividad del *tagger*. Con respecto al conjunto de etiquetas la granularidad es la característica que más afecta —al estar directamente relacionada con el tamaño—.

Si la granularidad del conjunto de etiquetas considerado es demasiado grande, la efectividad del *tagger* será muy alta, ya que sólo se consideran las distinciones más importantes y como consecuencia los datos pueden ser muy pobres. Por el contrario, si la granularidad del conjunto de etiquetas es muy fina, la efectividad del *tagger* será mucho menor. Además muchas distinciones finas no pueden resolverse sólo con información sintáctica o contextual, necesitándose conocimiento semántico.

En el campo de los corpora podemos considerar que existen los "corpora canónicos" que son aquellos que existen hace algunos años: Brown, Lancaster-Oslo-Bergen (LOB) y London-Lund. Los más conocidos son probablemente el Brown Corpus y el LOB. El Brown contiene un millón de palabras del inglés americano y fue anotado en 1979 empleando el *tagger* TAGGIT. El LOB etiquetado el mismo año y siguiendo los mismos criterios tiene el mismo número de palabras.

Actualmente, los corpora tienden a ser más grandes, y son compilados principalmente a

[9]Ver [Atkins et al., 1992] para futura información sobre diseño y desarrollo de un corpus.

través de proyectos e iniciativas tales como LDC, CLR —Consortium for Lexical Research—, EDR —Electronic Dictionary Research—, ECI o ACL/DCI. Estas asociaciones proveen corpora como el Wall Street Journal (WSJ, 300 millones de palabras de inglés americano), Lancaster Spoken English Corpus (SEC), Nijmegen TOSCA corpus, Bank of English corpus (BoE —con 200 millones de palabras—, y etiquetado con el entorno ENGCG [Järvinen, 1994]), o los 100 millones de palabras de British National Corpus (BNC) etiquetado con CLAWS *tagger* [Leech et al., 1994].

Aunque muchos corpora limitan su nivel de anotación a etiquetas POS (Part-Of-Speech), algunos ofrecen un alto nivel de anotaciones y constituyen una fuente importante de conocimiento para esta investigación. Podemos encontrar corpora analizados sintácticamente tales como el Susanne corpus, el Penn Treebank (3 millones de palabras) [Marcus et al., 1993] o el IBM/Lancaster treebank.

Asimismo, como ejemplo de corpus anotado a varios niveles (sintáctica y semánticamente) podemos considerar SemCor [Miller et al., 1993] que trataremos a continuación por ser el utilizado en nuestros experimentos (ver Capítulo 5), por su disponibilidad, así como por ser el más utilizado entre los investigadores.

Hasta hace pocos años, los corpora existentes estaban en lengua inglesa. Sin embargo, el éxito y la aplicabilidad de los corpora lingüísticos al PLN, han propiciado una rápida extensión a otras lenguas. Por ejemplo, Tresor de la Lange Française (TLF), que contiene 150 millones de palabras escritas en francés, o el corpus LEXESP que contiene 5 millones de textos equilibrados y representativos en español (ver figura 3.1). Una buena fuente de información sobre recursos léxicos en español es el Instituto Cervantes [Instituto-Cervantes, 1999].

Asimismo, un recurso ampliamente utilizado en traducción automática son los corpora paralelos. Un corpus paralelo se refiere a una colección o conjunto de textos, donde cada uno de éstos es la traducción del original a otras lenguas.

El caso más simple es cuando se trata sólo de dos lenguas o idiomas, así aparece el corpus bilingüe que consta a su vez de dos corpora, cada uno en una lengua, donde uno es la traducción exacta del otro. Uno de los mejores corpora bilingües conocidos es el Canadian Hansard, que consta de 500 millones de transcripciones en francés e inglés de las actas del parlamento canadiense.

Sin embargo, algunos corpora paralelos existen en varias lenguas, un ejemplo de este tipo es "La Biblia Políglota" [Davies, 1999] (ver figura 3.2) en el que el evangelio de San Lucas aparece escrito en trece lenguas: griego, latín, español antiguo, español moderno, portugués, francés, italiano, rumano, alemán, inglés antiguo, medieval y moderno e inglés actual.

La finalidad de los corpora paralelos es ayudar a la comunicación de sociedades u organizaciones multilingües, tales como Naciones Unidas, OTAN, así como a países oficialmente bilingües como Canadá.

```
MEDARDO_Fraile medardo_fraile NP00000
juega jugar VMIP3S0
a a SPS00
un un TIMS0
cinismo cinismo NCMS000
fácil fácil AQ0CS00
y y CC00
divertido divertido AQ0MS00
. . Fp
No no RG000
quiero querer VMIP1S0
decir decir VMN0000
que que CS00
lo ello PP3CS000
sea ser VASP3S0 ser VASP1S0
, , Fc
Cínico cínico AQ0MS00
o o CC00
divertido divertido AQ0MS00
, , Fc
sino_que sino_que CS00
ante ante SPS00
un un TIMS0
mazo mazo NCMS000
de de SPS00
hojas hoja NCFP000
grabadas grabar VMPP0PF
coloca colocar VMMP2S0
un un MCMS00
cristal cristal NCMS000
bien bien RG000
tallado tallado AQ0MS00
y …
```

Figura 3.1: Fragmento del corpus LEXESP

SemCor

SemCor es un subconjunto del Brown Corpus, etiquetado con información semántica por el mismo equipo que diseñó WordNet [Miller et al., 1993]. Todos los nombres, verbos, adjetivos y adverbios están etiquetados con el sentido correspondiente a la base de datos léxica WordNet. Miller et al. [1993] define a SemCor (Semantic Concordance) como:

"a textual corpus and lexicon so combined that every substantive word in the text is linked to its appropriate sense in the lexicon".

SemCor se ha construido a partir de dos corpora de textos: por un lado, incluye 103 pasajes del corpus estándar Present-Day editado por American English (el Brown Corpus) y por otro, consta del texto completo de la novela de Stephen Crane *The Red Badge of Courage*. El lexicón utilizado es la base de datos léxica WordNet.

SemCor consta de 500 pasajes de 2.000 palabras cada uno, extraídos de ediciones de documentos contemporáneos. Fue diseñado como una colección de textos heterogénea y equilibrada

The Polyglot Bible

Mark Davies, Illinois State University

CH:V	Spanish_1900s	English_1900s	French
6:1	Aconteció que Jesús pasaba por los sembrados en sábado, y sus discípulos arrancaban espigas y las comían, restregándolas con las manos.	One Sabbath Jesus was going through the grainfields, and his disciples began to pick some heads of grain, rub them in their hands and eat the kernels.	Il arriva, un jour de sabbat appelé second-premier, que Jésus traversait des champs de blé. Ses disciples arrachaient des épis et les mangeaient, après les avoir froissés dans leurs mains.
6:2	Y algunos de los fariseos dijeron: --¿Por qué hacéis lo que no es lícito hacer en los sábados?	Some of the Pharisees asked, "Why are you doing what is unlawful on the Sabbath?"	Quelques pharisiens leur dirent: Pourquoi faites-vous ce qu'il n'est pas permis de faire pendant le sabbat?
6:3	Respondiéndoles, Jesús dijo: --¿No habéis leído qué hizo David cuando tuvo hambre él y también los que estaban con él?	Jesus answered them, "Have you never read what David did when he and his companions were hungry?	Jésus leur répondit: N'avez-vous pas lu ce que fit David, lorsqu'il eut faim, lui et ceux qui étaient avec lui;
6:4	Entró en la casa de Dios, tomó los panes de la Presencia, que no es lícito comer, sino sólo a los sacerdotes, y comió y dio también a los que estaban con él.	He entered the house of God, and taking the consecrated bread, he ate what is lawful only for priests to eat. And he also gave some to his companions."	comment il entra dans la maison de Dieu, prit les pains de proposition, en mangea, et en donna à ceux qui étaient avec lui, bien qu'il ne soit permis qu'aux sacrificateurs de les manger?
6:5	--También les decía--: El Hijo del Hombre es Señor del sábado.	Then Jesus said to them, "The Son of Man is Lord of the Sabbath."	Et il leur dit: Le Fils de l'homme est maître même du sabbat.
6:6	Aconteció en otro sábado que él entró en la sinagoga y enseñaba. Y estaba allí un hombre cuya mano derecha estaba paralizada.	On another Sabbath he went into the synagogue and was teaching, and a man was there whose right hand was shriveled.	Il arriva, un autre jour de sabbat, que Jésus entra dans la synagogue, et qu'il enseignait. Il s'y trouvait un homme dont la main droite était sèche.
6:7	Los escribas y los fariseos le acechaban para ver si le sanaría en sábado, para hallar de qué acusarle.	The Pharisees and the teachers of the law were looking for a reason to accuse Jesus, so they watched him closely to see if he would heal on the Sabbath.	Les scribes et les pharisiens observaient Jésus, pour voir s'il ferait une guérison le jour du sabbat: c'était afin d'avoir sujet de l'accuser.
6:8	Pero él, conociendo los razonamientos de ellos, dijo al hombre que tenía la mano paralizada: --Levántate y ponte en medio. El se levantó y se puso en medio.	But Jesus knew what they were thinking and said to the man with the shriveled hand, "Get up and stand in front of everyone." So he got up and stood there.	Mais il connaissait leurs pensées, et il dit à l'homme qui avait la main sèche: Lève-toi, et tiens- toi là au milieu. Il se leva, et se tint debout.
6:9	Entonces Jesús les dijo: --Yo os pregunto: ¿Es lícito en el sábado hacer bien o hacer mal? ¿Salvar la vida o quitarla?	Then Jesus said to them, "I ask you, which is lawful on the Sabbath: to do good or to do evil, to save life or to destroy it?"	Et Jésus leur dit: Je vous demande s'il est permis, le jour du sabbat, de faire du bien ou de faire du mal, de sauver une personne ou de la tuer.
6:10	Y mirándolos a todos en derredor, dijo al hombre: --Extiende tu mano. El lo hizo, y su mano le fue restaurada.	He looked around at them all, and then said to the man, "Stretch out your hand." He did so, and his hand was completely restored.	Alors, promenant ses regards sur eux tous, il dit à l'homme: Étends ta main. Il le fit, et sa main fut guérie.

Figura 3.2: Corpus paralelo de la Biblia Políglota

a través de diferentes estilos y géneros literarios, tratando temas políticos, científicos, literarios, deportivos, musicales, cinematográficos, etc.

En la figura 3.3 se muestra un fragmento de un texto de SemCor, correspondiente a la frase siguiente del Brown Corpus:

The Fulton County Grand Jury said Friday an investigation of Atlanta's recent primary election produced "no evidence" that any irregularities took place.

Ésta se representa etiquetada en SGML, pudiéndose observar las etiquetas empleadas para indicar el sentido de cada uno de los términos que componen la frase. Así por ejemplo, el término *investigation* aparece etiquetado con el sentido 1 de WordNet (wnsn=1).

El corpus de textos que consta de la novela de Stephen Crane se compone de 24 capítulos cortos sobre la guerra civil americana con un total de 45.600 palabras.

Mientras que el Brown Corpus [Francis, 1982] contiene 1 millón de palabras del inglés americano, SemCor[10] consta de 250.000 palabras, donde todas ellas se etiquetaron manualmente con los sentidos de WordNet.

En la tabla 3.1 se muestra la composición de la versión 1.6 de SemCor.

La construcción de SemCor fue manual empleando como ayuda herramientas software diseñadas especialmente para este propósito[11]. Además, el proceso de anotación de SemCor fue

[10] Ver estadísticas en el Apéndice A.

[11] Se utilizó ConText (programa etiquetador semántico) para etiquetar la prosa con los sentidos de WordNet.

```
<contextfile concordance=brown>
<context filename=br-a01 paras=yes>
<p pnum=1>
<s snum=1>
<wf cmd=ignore pos=DT>The</wf>
<wf cmd=done rdf=group pos=NNP lemma=group wnsn=1
lexsn=1:03:00:: pn=group>Fulton_County_Grand_Jury</wf>
<wf cmd=done pos=VB lemma=say wnsn=1
lexsn=2:32:00::>said</wf>
<wf cmd=done pos=NN lemma=friday wnsn=1
lexsn=1:28:00::>Friday</wf>
<wf cmd=ignore pos=DT>an</wf>
<wf cmd=done pos=NN lemma=investigation wnsn=1
lexsn=1:09:00::>investigation</wf>
<wf cmd=ignore pos=IN>of</wf>
<wf cmd=done pos=NN lemma=atlanta wnsn=1
lexsn=1:15:00::>Atlanta</wf>
<wf cmd=ignore pos=POS>'s</wf>
<wf cmd=done pos=JJ lemma=recent wnsn=2
lexsn=5:00:00:past:00>recent</wf>
<wf cmd=done pos=NN lemma=primary_election wnsn=1
lexsn=1:04:00::>primary_election</wf>
<wf cmd=done pos=VB lemma=produce wnsn=4
lexsn=2:39:01::>produced</wf>
<punc>``</punc>
<wf cmd=ignore pos=DT>no</wf>
<wf cmd=done pos=NN lemma=evidence wnsn=1
lexsn=1:09:00::>evidence</wf>
<punc>''</punc>
<wf cmd=ignore pos=IN>that</wf>
<wf cmd=ignore pos=DT>any</wf>
<wf cmd=done pos=NN lemma=irregularity wnsn=1
lexsn=1:04:00::>irregularities</wf>
<wf cmd=done pos=VB lemma=take_place wnsn=1
lexsn=2:30:00::>took_place</wf>
<punc>.</punc>
</s>
</P>
```

Figura 3.3: Fragmento de un texto de SemCor

utilizado por los lexicógrafos como un mecanismo de mejorar la cobertura de WordNet acordemente, pues se detectaron palabras y significados perdidos, así como sentidos no distinguibles.

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 o significados de palabras polisémicas.

En el Apéndice A se incluyen unas estadísticas de la colección SemCor, la estructura y descripción en SGML de los documentos, así como las etiquetas sintácticas utilizadas.

SemCor		Contenido
Brown1	103	ficheros Brown Corpus
Brown2	83	ficheros Brown Corpus
Brownv	166	ficheros Brown Corpus (Verbos)

Tabla 3.1: Contenido de SemCor 1.6

Conjunto categorías	Número categorías	No categorías 1+ocurrencia	No categorías 20+ocurrencias
Topics	135	120	57
Organizations	56	32	9
Exchanges	39	32	7
Places	175	147	60
People	267	114	15

Tabla 3.2: Clasificación de categorías Reuters-21578

Reuters

La colección de documentos Reuters [Lewis, 1992] es un conjunto de partes periodísticos aparecidos en el canal de noticias Reuters durante el año 1987, ensamblados y categorizados manualmente por personal de Reuters Ltd. y del Carnegie Group, Inc. Los documentos de la colección Reuters son noticias de carácter económico, cuyo tamaño varía desde una línea hasta más de una página. Esta colección ha sido de gran ayuda para la tarea de categorización de textos [Buenaga et al., 1997], así como para la recuperación de información [Sanderson, 1996], debido a su carácter estándar y a su libre distribución.

Los documentos de la colección Reuters fueron clasificados dentro de un conjunto de 672 categorías (ver tabla 3.2) divididas en *Topics* (135), *Organizations* (56), *Exchanges* (39), *Places* (175) y *People* (267).

El primer conjunto (Topics) contiene categorías de carácter económico, basadas en el contenido semántico del texto, y son las más relevantes a esta investigación. Ejemplos de estas categorías incluyen "Conut", "Gold", "Inventories" y "Money-supply" entre otras. Este conjunto de categorías ha sido el más utilizado en investigación. Existen 10.509 documentos clasificados como no pertenecientes a ninguna de las categorías de Topics. Mientras que los últimos cuatro conjuntos de categorías (Exchanges, Orgs, People y Places) se refieren a organizaciones, lugares, etc. importantes en el ámbito económico durante 1987, y se corresponden con entidades específicas. Ejemplos de ellas incluyen "nasdaq" (Exchange), "gat" (Orgs), "perez-de-cuellar" (People), y "spain" (Places). Un documento típico asignado a una categoría explícitamente incluye de alguna forma el nombre de la categoría en los documentos de textos. Sin embargo, no todos los documentos que contengan explícitamente el nombre de una determinada categoría son asignados a dicha categoría, pues depende del enfoque de la noticia [Hayes y Weinstein, 1990].

Así, estas categorías genéricas no son fáciles de asignar como cabría pensar.

En la tabla 3.2 se muestran cuantas categorías aparecen al menos en uno de los 21.578 documentos de la colección, y cuantas aparecen al menos en veinte de los documentos [Lewis, 1997] [12].

```
<REUTERS TOPICS=''YES'' LEWISSPLIT=''TEST''
CGISPLIT=''TRAINING-SET'' OLDID=''6505'' NEWID=''18753''>
<DATE>18-JUN-1987 11:44:27.20</DATE>
<TOPICS><D>bop</D><D>trade</D></TOPICS>
<PLACES><D>italy</D></PLACES> <TEXT> <TITLE>ITALIAN BALANCE OF
PAYMENTS IN DEFICIT IN MAY </TITLE> <BODY>
 Italy's overall balance of payments showed a deficit
 of 3,211 billion lire in May compared with a surplus
 of 2,040 billion in April, provisional Bank of Italy
 figures show. The May deficit compares with a surplus
 of 1,555 billion lire in the corresponding month of
 1986. For the first five months of 1987, the overall
 balance of payments showed a surplus of 299 billion
 lire against a deficit of 2,854 billion in the
 corresponding 1986 period. REUTER
</BODY> </TEXT>
```

Figura 3.4: Documento 18.753 de Reuters-21.578

En la figura 3.4 puede observarse un ejemplo de documento correspondiente a la revisión 21.578 anotado en SGML.

En el Apéndice D se incluye información sobre Reuters.

3.3 Lexicón

El conocimiento léxico —conocimiento sobre palabras individuales en el lenguaje— es esencial para todo tipo de procesamiento del lenguaje natural. El lexicón puede contener una amplia gama de información sobre palabras específicas, dependiendo de la estructura y de la tarea del sistema de procesamiento del lenguaje natural. Se hace necesario para sistemas que analicen el contenido de textos reales la existencia de lexicones más grandes y más ricos [Walker et al., 1995].

Un paso importante y crítico, para evitar esfuerzos duplicados y consecuentemente hacia una mayor productividad, es la construcción de recursos a escala real. Recursos léxicos de amplia cobertura con tipos básicos de información, que estén públicamente disponibles a la comunidad y lo suficientemente genéricos para poder ser reutilizados en diferentes aplicaciones. Esta nece-

[12] De hecho, ocurre que algunas categorías no aparecen en ningún documento de la colección.

sidad de compartir recursos, construidos posiblemente en cooperación, conlleva la agregación de especificaciones comunes y consensuadas hacia una estandarización [Zampolli et al., 1994].

Un lexicón básico incluirá información morfológica, bien por medio de la generación de todas las formas de palabras asociadas con características pertinentemente morfosintácticas, o mediante una lista de formas de palabras, o así como una combinación de ambas. El nivel sintáctico incluirá las estructuras de cada palabra o el significado de la palabra. Por otro lado, un lexicón más complejo puede incluir información semántica, como una clasificación jerárquica. Para la tarea de traducción automática el lexicón tendrá que registrar la correspondencia entre los elementos léxicos, tanto en el lenguaje origen como en el de destino. Para la comprensión del lenguaje y generación, éste tendrá que incluir información sobre la pronunciación de palabras individuales (transcripciones fonéticas).

Tradicionalmente, los lexicones computerizados se han construido específicamente para el análisis y la generación del lenguaje. Estos lexicones han sido construidos a mano (costosos) para satisfacer las necesidades de sistemas individuales, donde no han sido tratados como recursos principales para ser compartidos por otros grupos. Trataremos primero las bases de datos léxicas y nos centraremos en WordNet, una base de datos léxica del inglés que ha sido la más utilizada en trabajos relacionados con WSD y TC.

3.3.1 Bases de datos léxicas

Con el término base de datos léxica se hace referencia en la bibliografía a diversos tipos de sistemas. En todos ellos, el propósito fundamental es el de almacenar información relativa a un conjunto de términos de una o más lenguas, aunque la cobertura puede variar de unos a otros, así como los diferentes tipos de información existente para cada término.

Durante 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 ejemplos de este tipo de proyectos podemos citar WordNet [Miller et al., 1993; Miller, 1995], EuroWordNet [Vossen y otros, 1998], el léxico incorporado en el sistema ALVEY [Carroll y Grover, 1993], el desarrollado en el proyecto ACQUILEX [Verdejo, 1994], y EDR [Yokoi, 1995]. Asimismo, no sólo se han centrado en el léxico, sino en el conocimiento enciclopédico como el proyecto CyC[13] que se propuso desarrollar una base de conocimiento de propósito general a escala real, con información de carácter "enciclopédico".

EuroWordNet es una base de datos léxica multilingüe (incluye varias lenguas de la Comunidad Europea). Este proyecto[14] se centra en la construcción de una base de datos con relaciones semánticas entre palabras para varias lenguas de la Comunidad Europea (alemán, italiano, español, francés, checo y estonio). Las redes de palabras serán enlazadas con WordNet (base

[13] Es un proyecto de la MCC (Microelectronics & Computer Technology Corporation).

[14] El proyecto EuroWordNet está financiado por el EC (LE2-4003) y un conjunto de instituciones: la Universidad de Amsterdam (coordinador), la Universidad de Sheffield, el Instituto Lingüística Computacional del CNR (Pisa), la Fundación Universidad-Empresa (una cooperación de UNED Madrid y la UPC Barcelona).

de datos léxica en inglés) compartiendo su ontología, mientras que las redes de palabras serán mantenidas en redes de palabras individuales.

El proyecto ACQUILEX tuvo como principal objetivo la reutilización de información extraída de diccionarios electrónicos. La adquisición de información sintáctica y semántica se realizó usando técnicas y metodologías comunes de extracción de información de más de diez diccionarios en cuatro lenguas. El objetivo fue construir una base de conocimiento léxico común mediante un prototipo, utilizando un único sistema tipo para todas las lenguas y diccionarios considerados.

EDR (Electronic Dictionary Research) fue un proyecto realizado en Japón incluye diccionarios en inglés y japonés y diccionarios bilingües inglés-japonés y japonés-inglés. Abarca más de 400.000 conceptos, conteniendo las palabras del diccionario información gramatical y enlaces al concepto jerárquico.

Como se ha expuesto los sistemas proporcionan información sobre uno, dos o más idiomas, lo que puede resultar determinante para operaciones sobre textos como la traducción automática. En concreto WordNet incluye información sólo para los términos del inglés, EDR incluye información para el inglés y el japonés, y ACQUILEX para cuatro lenguas de la Comunidad Europea.

También varían los tipos de información incluidos en cada sistema de unos casos a otros. Por ejemplo, en ACQUILEX se almacena información para un número mayor de relaciones léxicas que en WordNet. En EDR se encuentran almacenados además, textos completos asociados a los diferentes significados de cada uno de los términos (adviértase el incremento en el tamaño absoluto de la base de datos que esto conlleva). En todos los casos, una cuestión determinante es la inclusión de información para un conjunto de términos próximo a la totalidad de los del idioma. La información debe ser además lo suficientemente rica y estructurada como para ser utilizada en tareas de procesamiento de textos basadas en su contenido.

Los proyectos también han diferido en su grado de completitud o terminación, dependiendo de la magnitud de la información que ha sido necesario incluir y otra serie de factores específicos en cada caso. En la actualidad, un sistema muy utilizado en sistemas de clasificación de documentos ha sido WordNet [Hearst, 1994; Buenaga, 1996] tanto por su completitud como por su disponibilidad, al ser de dominio público.

Históricamente, las bases de datos léxicas han tenido un especial interés desde el punto de vista del desarrollo de sistemas de Procesamiento del Lenguaje Natural [Allen, 1994] y los sistemas de Recuperación de Información [Amsler, 1989; Salton, 1989].

Los sistemas enmarcados en el campo del Procesamiento de Lenguaje Natural (por ejemplo interfaces en lenguaje natural, sistemas de traducción automática, de extracción de información y generación de resúmenes), normalmente necesitan una gran cantidad de información léxica y gramatical, cuya confección conlleva un importante esfuerzo de desarrollo. Desde este punto de vista, disponer de una base de datos léxica puede disminuir en gran medida el desarrollo de este tipo de sistemas, al permitir obtener gran parte del léxico del sistema de forma automática [Guthrie, 1996].

Desde el punto de vista de la resolución de la ambigüedad léxica y de los sistemas de clasi-

ficación de documentos, la utilización de bases de datos léxicas puede permitir un análisis más rico de los textos a procesar [Ureña et al., 1998c; Buenaga et al., 1997].

Interesa relacionar las bases de datos léxicas con los diccionarios electrónicos. Dentro de una escala de dimensiones y de complejidad estructural, las bases de datos léxicas se encontrarían en un punto superior a los diccionarios electrónicos. Las bases de datos léxicas parecen marcar el grado máximo al que la tecnología actual es capaz de llegar, conservando la deseada generalidad de su cobertura.

Prestaremos atención especial a WordNet, por constituir un recurso lingüístico base para nuestra investigación en WSD.

WordNet

WordNet [Miller, 1990; Miller et al., 1993; Miller, 1995; Fellbaum, 1998] es un sistema que posee información léxica extraída de forma semi-automática de diccionarios. Este sistema ha sido desarrollado en el Cognitive Science Laboratory de la Universidad de Princeton.

WordNet tiene su origen en un proyecto cuyo objetivo era producir un diccionario que permitiera búsquedas conceptuales en lugar de alfabéticas, inspirándose su diseño en teorías psicolingüísticas sobre la memoria léxica humana. Con el paso del tiempo, WordNet se ha convertido en una base de datos de más de 15 megabytes (versión 1.5) con información sobre términos pertenecientes a las cuatro categorías sintácticas más importantes: nombres, verbos, adjetivos y adverbios[15]. Una muestra de la información existente en WordNet se incluye en la figura 3.5[16].

En WordNet, el elemento básico que permite representar conceptos como conjunto de sinónimos es el denominado *synset*.

En WordNet también se almacena información relativa a las diferentes relaciones definidas entre palabras y *synsets* (o conceptos). Trataremos a continuación las más importantes: hiponimia, meronimia y antonimia, a parte de la misma sinonimia, de la que se hace uso implícito en la definición de los *synsets*.

En los puntos siguientes tratamos con detalle los aspectos más importantes de WordNet con vistas a su utilización en sistemas de desambiguación.

Relaciones en WordNet

WordNet hace uso de una distinción comúnmente aceptada entre las relaciones conceptuales-semánticas, que enlazan conceptos y las relaciones léxicas que relacionan palabras individuales. La información almacenada en WordNet para cada término viene dada fundamentalmente por las relaciones léxicas en las que interviene con otros términos. La red (semántica) que es definida por estas relaciones existentes entre las palabras, sirve para dar nombre al sistema. En la figura

[15] La versión 1.5 contiene más de 126.000 entradas, de las cuales el 70% son nombres, el 15% adjetivos, el 10% verbos y el 5% restante son adverbios.

[16] Extraída de [Vaquero y Buenaga, 1996]

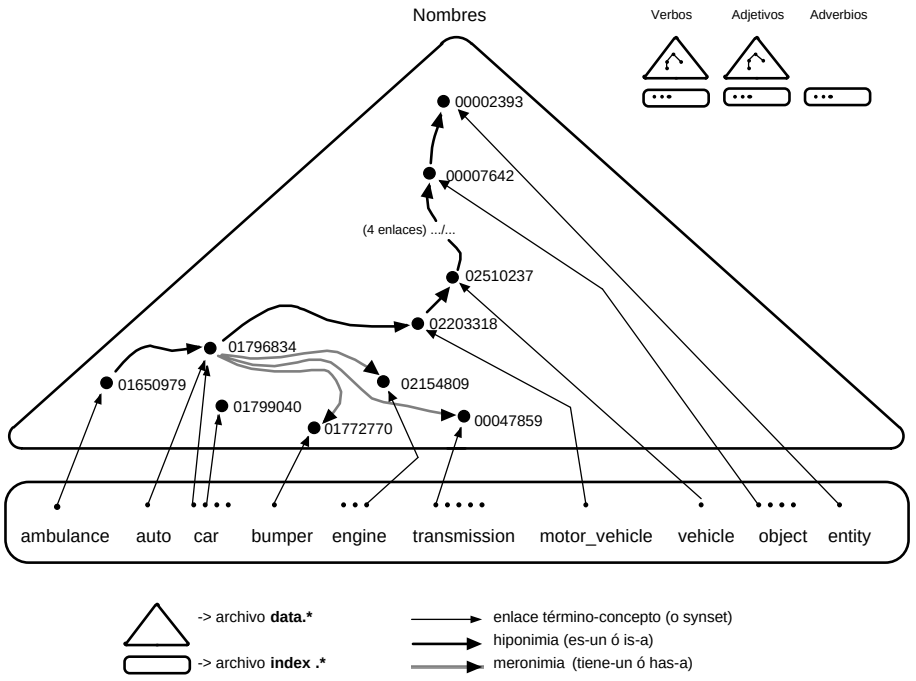

Figura 3.5: Estructura y relaciones de WordNet

3.6 se presenta un ejemplo de la información que existe en la base de datos, asociada al término "plant". Tratamos a continuación las principales relaciones léxicas existentes en WordNet.

La relación de sinonimia representa en WordNet un papel determinante. Dos palabras son sinónimas cuando tienen un significado común. La relación de sinonimia se utiliza como base para la definición del objeto básico en WordNet: el *synset*.

En WordNet, por definición, un *synset* es un conjunto de sinónimos estrictos. A su vez, cada *synset* en que aparece una palabra representa un significado diferente de la palabra. Los *synsets* {board, plank} y {board, committee} pueden servir como designadores no ambiguos de los dos significados de "board". De forma intuitiva podemos hacer corresponder un *synset* con un concepto. Una palabra puede referenciar diferentes conceptos, o *synsets*. A su vez, un mismo concepto puede ser referenciado por varias palabras. En la figura 3.6 puede verse como el término "plant" se encuentra asociado a 4 *synsets* o conceptos, que dan cuenta de cada de sus posibles significados en inglés. El *synset* sirve como objeto básico en torno al cual se definen las restantes relaciones léxicas en WordNet.

La hiponimia (hyponymy) es la relación existente entre conceptos o *synsets* equivalente a la relación de generalización, o *is-a*, o *es-un*. Un concepto representado por el *synset* {x, x', ...} es

```
4 senses of plant
Sense 1
plant, works, industrial plant
 => building complex, complex
Sense 2
plant, flora, plant life
 => life form, organism, being,
    living thing
Sense 3
plant
 => contrivance, stratagem, dodge
Sense 4
plant
 => actor, histrion, player,
    thespian,
    role player
```

a) synsets de "plant"

```
Sense 2
plant, flora, plant life
 => life form, organism, being,
    living thing
   => entity, something
```

b) hiperónimos del synset #2

```
Sense 2
plant, flora, plant life
 HAS SUBSTANCE: plant tissue
 HAS PART: plant part
 => life form, organism, being,
    living thing
 HAS SUBSTANCE: tissue
 HAS PART: cell
   HAS PART: energid, protoplast
   HAS PART: cytoplasm
     HAS PART: microsome
     HAS PART: Golgi body, Golgi
     apparatus,
     Golgi complex
   HAS PART: nucleus, cell nucleus
     HAS PART: chromatin
     HAS PART: achromatin
     HAS PART: chromosome
   ../..
   HAS PART: organelle, cell organ
   HAS PART: vacuole
   HAS PART: cell wall, cell
   membrane, plasma membrane
 HAS PART: body part
   HAS PART: corpus
```

c) merónimos del synset #2

```
Sense 2
plant, flora, plant life
       => phytoplankton
       => ornamental
       => acrogen
       => apomict
       => aquatic
       => cryptogam
       => annual
       => biennial
       => perennial
       => escape
       => embryo
       => monocarp, monocarpic plant,
          monocarpous plant
       => sporophyte
       => gametophyte
       => fungus
       => houseplant
       => garden plant
       => vascular plant, tracheophyte
       => poisonous plant
       => air plant, epiphyte,
          aerophyte, epiphytic plant
       ../..
       => myrmecophyte
```

d) hipónimos del synset #2

Figura 3.6: Muestras de información asociada en WordNet

un hipónimo del concepto representado por el *synset* {y, y', ..}, si un hablante del inglés acepta oraciones construidas de la forma "an x is (a kind of) y". Por ejemplo, {maple} es un hipónimo de {tree} y {tree} es un hipónimo de {plant}. La hiperonimia (hypernymy) se define como la relación inversa a la hiponimia (figura 3.6).

La relación de meronimia (meronymy) da cuenta de la inversa de *tiene-un* o *has-a*. Un concepto representado por el *synset* {x, x', ...} es un merónimo del concepto representado por el *synset* {y, y', ...}, si un hablante del inglés acepta oraciones construidas de la forma "a y

has an x" , o "an x is a part of y". Por ejemplo, para el *synset* de "plant" {plant, flora, plant life}, se encuentran como partes constituyentes *synsets* correspondientes a partes de una planta, cuerpo, tallo, etc. (figura 3.6). La holonimia (holonymy) se define como la relación inversa a la meronimia.

A diferencia de la hiponimia y la meronimia, la relación de antonimia (antonymy) en WordNet se encuentra a nivel de palabras, no de conceptos. Por ejemplo, los significados {rise, ascend} y {fall, descend} son conceptualmente diferentes, pero no son antónimos, sin embargo sí son antónimos los términos individualmente [rise/fall] y [ascend/descend].

La información de meronimia e hiponimia existente en WordNet es equivalente a la de las jerarquías definidas en redes semánticas mediante la relación *es-un* y *tiene-un*, con la importante característica de sus dimensiones: los 95.600 términos que contiene WordNet en la actualidad, se encuentran organizados en 70.100 *synsets*.

En WordNet se incluye también información sintáctica. Para cada palabra se incluyen las categorías sintácticas que puede tener. De hecho, esta información se utiliza como base para la estructuración de los conceptos. En concreto, el léxico se divide en cuatro categorías: nombres, verbos, adjetivos y adverbios, de una forma análoga a los diccionarios comunes. De esta forma, los conceptos se encuentran agrupados en cuatro conjuntos disjuntos correspondientes a estas categorías. La mayor parte de la información almacenada corresponde a los nombres y los verbos. Para cada categoría se definen diferentes relaciones semánticas, además de las anteriormente citadas. Por otra parte, la meronimia sólo se incluye para los nombres, y para los adjetivos y adverbios no se define la hiponimia.

WordNet ha sido empleada en variados trabajos relacionados con desambiguación y clasificación de textos, debido a su amplia cobertura (126.520 palabras). Su uso principal se refiere a la tarea de desambiguación de significados. Ello se debe a que en este problema es preciso definir de alguna manera precisa el significado de una palabra, y la definición basada en conjuntos de sinónimos es precisa y se halla en formato electrónico. WordNet también ha sido usada en categorización de textos, al permitir establecer relaciones entre términos y categorías de manera automática sin recurrir a un conjunto de documentos previamente categorizados.

Finalmente, apuntar que WordNet es muy popular en el PLN y, está disponible en Internet[17]. En el Apéndice B se incluye la estructura de WordNet.

3.3.2 Diccionarios electrónicos

Entre los primeros lexicones construidos como recursos compartidos para la lingüística computacional se encuentran los diccionarios electrónicos. Mediante este término, o diccionario en formato electrónico *(machine readable dictionary)* se suele hacer referencia a diccionarios ya existentes (publicados en formato físico habitual) disponibles en forma electrónica [Guthrie, 1996].

La diferencia que se puede establecer entre los diccionarios electrónicos y las bases de datos léxicas reside en los diferentes tipos de información que incluyen y su grado de estructuración.

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

Una base de datos léxica se encuentra más próxima a su utilización directa desde un sistema de procesamiento de lenguaje natural o de clasificación de documentos que un diccionario electrónico.

Desde el punto de vista de las bases de datos léxicas, los diccionarios electrónicos han resultado de especial interés por permitir la construcción de la base de datos léxica de forma semiautomática. Esto ocurre especialmente cuando en la definición del formato del diccionario los editores prestan especial interés a la estructuración de aspectos sintácticos y de los diferentes significados de las palabras.

```
Searching for 'car'

Vehicle. 272
       post chaise, diligence, stage; stage coach, mail coach,
       hackney coach, glass coach; stage wagon, car, omnibus, fly,
       cabriolet[obs3], cab, hansom, shofle[obs3], four-wheeler,
       growler, droshki[obs3], drosky[obs3].

Vehicle. 272
       motor car, automobile, limousine, car, auto, jalopy,
       clunker, lemon, flivver, coupe, sedan, two-door sedan,
       four-door sedan, luxury sedan; wheels [coll.], sports car,
       roadster, gran turismo[It], jeep, four-wheel drive vehicle,
       electric car, steamer; golf cart, electric wagon; taxicab,
       cab, taxicoach[obs3], checker cab, yellow cab; station
       wagon, family car; motorcycle, motor bike, side car;
       van, minivan, bus, minibus, microbus; truck, wagon, pick-up
       wagon, pick-up, tractor-trailer, road train, articulated
       vehicle; racing car, racer, hot rod, stock car, souped-up
       car..

Vehicle. 272
       post chaise, diligence, stage; stage coach, mail coach,
       hackney coach, glass coach; stage wagon, car, omnibus, fly,
       cabriolet[obs3], cab, hansom, shofle[obs3], four-wheeler,
       growler, droshki[obs3], drosky[obs3].

Vehicle. 272
       motor car, automobile, limousine, car, auto, jalopy,
       clunker, lemon, flivver, coupe, sedan, two-door sedan,
       four-door sedan, luxury sedan; wheels [coll.], sports car,
       roadster, gran turismo[It], jeep, four-wheel drive vehicle,
       electric car, steamer; golf cart, electric wagon; taxicab,
       cab, taxicoach[obs3], checker cab, yellow cab; station
       wagon, family car; motorcycle, motor bike, side car; van,
       minivan, bus, minibus, microbus; truck, wagon, pick-up
       wagon, pick-up, tractor-trailer, road train, articulated
       vehicle; racing car, racer, hot rod, stock car,
       souped-up car..
```

Figura 3.7: Muestra del thesaurus Roget con información relativa al término “car”

Los primeros esfuerzos realizados para la construcción de diccionarios electrónicos se encuentran en el Merriam-Webster’s 7th Collegiate Dictionary[18] y en el British dictionaries of English

[18]Primer diccionario electrónico disponible libremente.

language learners, utilizados experimentalmente en algunos sistemas. Entre los diccionarios electrónicos que han sido ampliamente utilizados en sistemas de lingüística computacional [Boguraev y Briscoe, 1989] podemos citar algunos como el Longman Dictionary of Contemporary English (LDOCE), el Oxford Advanced Learner's Dictionary of Current English (OALD) y el thesaurus Roget (figura 3.7).

Se ha puesto de manifiesto, a través de variadas aplicaciones del PLN, la necesidad de dotar a los diccionarios de una funcionalidad más completa [Agirre et al., 1995]. Entre las funciones que se hacen necesarias cabe destacar:

La búsqueda conceptual. Es una utilidad muy relevante en la generación léxica. Facilita la búsqueda de formas léxicas a partir de ideas y de conceptos abstractos.

La presentación de ejemplos de uso. De esta manera se muestra el manejo contextualizado de unidades léxicas.

La búsqueda de coocurrencias léxicas. Se trata de mostrar al usuario combinaciones de unidades léxicas que aparecen estrechamente ligadas, no tanto por criterios sintácticos o semánticos, sino de frecuencia de uso.

La incorporación de funciones gramaticales. La idea de intregrar analizadores y generadores morfosintácticos y otras herramientas para el tratamiento automático del lenguaje natural.

3.3.3 Thesaurus

El thesaurus es uno de los recursos lingüísticos candidato para ser utilizado por un desambiguador automático. Así numerosos investigadores han utilizado este recurso en la investigación de la desambiguación.

```
413 LONGITUDINAL
    TRANSVERSE

414 CRYOGENIC
    CRYOTRON
    PERSISTENT-CURRENT
    SUPERCONDUCT

415 ANTENNA
    KLYSTRON
    RECEIVER
    TRANSMITTER
    WAVEGUIDE
```

Figura 3.8: Fragmentos de thesaurus

Un thesaurus proporciona una agrupación o clasificación de términos en un determinado dominio o área en categorías denominadas clases. En la figura 3.8 aparece un ejemplo de thesaurus [Salton y McGill, 1983]. Pueden utilizarse diversos tipos de thesaurus en la indexación

automática o manual de los documentos [Srinivasan, 1992]. En una aproximación basada en el análisis automático de textos y consultas, la utilización del thesaurus permite identificar términos lexicográficamente diferentes como equivalentes semánticamente (por ejemplo sinónimos). La utilización de thesaurus se puede realizar durante el proceso de indexación de los documentos (utilizando como términos de indexación las clases del thesaurus), o durante el análisis de las consultas (realizándose una expansión de los términos de la consulta mediante sus equivalentes) [Srinivasan, 1992; Qiu y Frei, 1993].

```
(fla,flow,blood,serum, secretium)
(change, increase, effect, response, pattern)
(day, hour, week, year, month, hr, time)
(rat, mouse, animal, dog, female, infant)
```

Figura 3.9: Ejemplo de lista de asociación de términos

La construcción manual de un thesaurus es un proceso costoso. Existen métodos de construcción de thesaurus semiautomáticos y completamente automáticos [Srinivasan, 1992]. Los métodos de construcción de *listas de asociación de términos*, equivalentes a las clases del thesaurus anteriormente comentadas, permiten su obtención a partir del análisis de los documentos. Las listas de asociación son conjuntos de palabras con significado parecido a efectos de recuperación (ver figura 3.9). La suposición fundamental en que se basa la construcción de estas listas [Grefenstette, 1992] es que:

"Similar terms appear in similar contexts"

La siguiente expresión 3.1 se puede utilizar para el cálculo de la similitud entre dos términos $term_j$ y $term_k$ en una colección de n documentos [Salton y McGill, 1983; Srinivasan, 1992]:

$$sim(term_j, term_k) = \sum_{i=1}^{n} wd_{ij} \cdot wd_{ik} \tag{3.1}$$

En donde los valores wd_{ij} y wd_{ik} son los pesos de los términos $term_j$ y $term_k$ en el documento i. De esta forma, términos que aparecen simultáneamente en un número importante de documentos resultan con valores de similitud altos, mientras que términos que no aparecen en los mismo contextos, resultan con valores de similitud bajos, o nulos. A modo de ejemplo, si las palabras "permission" y "authorization" aparecen a la vez en un gran número de documentos, obtendremos una similitud para ellas importante. La formación de las asociaciones, o grupos, de términos se realiza a partir de sus valores de similitud pudiéndose utilizar diferentes algoritmos de agrupación (clustering) [Salton y McGill, 1983; Srinivasan, 1992].

3.4 Resumen y conclusiones

En este capítulo hemos hecho una presentación y clasificación de los recursos lingüísticos disponibles más relevantes para nuestro estudio, y que pueden mejorar los sistemas de procesamiento

de lenguaje natural. Nos hemos centrado principalmente en los corpora de textos y las bases de datos léxicas, puesto que son los recursos utilizados en el centro de esta investigación que es la resolución de la ambigüedad léxica, así como en otros recursos necesarios para las tareas de clasificación automática de documentos. El objetivo fundamental es incorporar recursos lingüísticos al proceso de desambiguador.

Capítulo 4

Resolución de la ambigüedad léxica

4.1 Introducción

La resolución automática de la ambigüedad léxica de términos polisémicos es una tarea muy útil, aunque compleja para muchas aplicaciones del procesamiento del lenguaje natural, como se ha tratado brevemente en el Capítulo 1. El objetivo principal de este capítulo es presentar un enfoque de desambiguación, así como los experimentos y evaluación directa de los resultados obtenidos.

Un algoritmo de desambiguación permite identificar el significado concreto con que aparece una palabra en un determinado contexto. Para desambiguar un término que aparece en un determinado lugar de un documento, utilizamos dos fuentes básicas de información. En primer lugar, los significados que tiene asociados y sus definiciones, para cada palabra tenemos asociados diferentes significados. Cada significado debe tener asociada una definición, en forma de texto o conjunto de palabras. Esta definición se puede obtener de un recurso lingüístico, como puede ser una base de datos léxica o un diccionario electrónico. En segundo, el contexto en que aparece el término, es decir las palabras que aparecen próximas a él en el texto del documento. Para desambiguar un término, esencialmente, calculamos la similitud entre el contexto en el que aparece y las definiciones de los significados. Para el cálculo de la similitud utilizamos el modelo del espacio vectorial y las técnicas de indexación vistas en el Capítulo 2. Se selecciona el significado que presenta mayor similitud con el contexto.

La organización de este capítulo es como sigue. Comenzamos describiendo la tarea WSD y su terminología. Seguidamente, se hace una breve revisión bibliográfica y una clasificación de los métodos de desambiguación, de acuerdo con el método utilizado en la adquisición de conocimiento. A continuación, se trata la efectividad del proceso de desambiguación, describiendo las métricas más utilizadas en su evaluación. Después, se presenta el sistema desambiguador y la metodología utilizada, y se proponen los enfoques de desambiguación, basado en corpus de textos, en bases de datos léxica y en la integración de ambos recursos lingüísticos. En un punto siguiente, se describe el entorno experimental, presentando una serie de experimentos orientados al estudio

de la efectividad de los enfoques propuestos de resolución de la ambigüedad. Finalmente, se presenta un resumen y conclusiones.

4.2 Descripción de la tarea y terminología

La tarea de un sistema desambiguador del significado es resolver la ambigüedad léxica de una palabra en un determinado contexto, tomando los sentidos proporcionados por un lexicón. Para precisar más, el término "ambigüedad léxica" hace referencia a dos conceptos "homonimia" y "polisemia". Se denomina homonimia cuando dos palabras comparten la misma forma léxica, y polisemia cuando una palabra posee varios significados. Convencionalmente, se ha venido utilizando como ejemplo de homonimia la distinción entre *bank* (finantial institution) y *bank* (river edge); y como ejemplo de polisemia *rust* (verbo) y *rust* (nombre). En esta disertación utilizaremos generalmente el término polisemia para referirnos a ambos tipos de ambigüedad léxica. Ya que la diferencia entre estos dos tipos, ha tenido menos controversias en tareas de desambiguación del significado de las palabras (aunque desde un punto de vista lingüístico ambos tipos de ambigüedad pueden ser rigurosamente definidos). Y porque el enfoque de esta investigación es la desambiguación tanto de nombres como de verbos. En los últimos años, se han propuesto diversos y variados sistemas de WSD. Así, la tarea WSD puede considerarse como una categorización, pues el significado plausible de una palabra puede seleccionarse de un conjunto de candidatos predefinidos. A continuación, vamos a comentar un poco de terminología.

4.2.1 Contexto

El contexto es la única manera de poder identificar el significado de palabras polisémicas. Todos los trabajos en desambiguación se basan en el contexto de la palabra dada para proporcionar información y ser usada en la desambiguación. El contexto se ha venido utilizando de dos maneras:

- Enfoque basado en *bolsa de palabras*: se considera como contexto las palabras de alrededor a la palabra dada, tomado como un grupo sin considerar las relaciones con la palabra dada en términos de distancia, ni las relaciones gramaticales [Schütze, 1998], etc.

- Enfoque basado en *información relacional*: el contexto se considera en términos de alguna relación a la palabra de que se trate, incluyendo distancia desde la palabra actual [Yarowsky, 1993, 1994a,b], las relaciones sintácticas [Earl, 1973], preferencias de selección [Hayes, 1977; Wilks, 1973; Hirst, 1987], propiedades ortográficas, colocaciones [Daghlgren, 1988; Atkins, 1987], categorías semánticas [Yarowsky, 1993], etc.

4.2.2 Sentidos o usos

La idea Aristotélica de que las palabras corresponden a objetos específicos así como a conceptos, fue reemplazada en este siglo por las ideas de Saussure y otros [Meillet, 1926; Hjemslev, 1953].

Por ejemplo, para Meillet el sentido de una palabra se define sólo por la media de sus usos lingüísticos. Un "uso" designa una ocurrencia o aparición de una palabra en un determinado contexto. Si tenemos dos frases que contienen una palabra o incluso en la misma frase en dos lugares diferentes, podemos decir que tenemos dos usos. Los usos por tanto son tipos. Wittgenstein toma una posición similar en su *Philosophische Utersuchungen*, afirmando que no hay sentidos, sino usos:

"For a large class of cases —though not for all— in which we employ the word 'meaning' it can be defined thus: the meaning of a word is its use in the language"

Puntos de vista similares aparecen en las últimas teorías sobre el significado. Para algunos, el significado es una función de distribución, mientras que para otros es una situación semántica, donde los sentidos de una palabra son considerados como una abstracción del papel que representa esto sistemáticamente en el discurso [Ide y Veronis, 1998].

El proyecto COBUILD [Sinclair, 1987] adopta este punto de vista de los significados, creando las divisiones de significados en el diccionario sobre la base de **clusters** de citaciones en un corpus.

Kilgarriff [1992] también implícitamente considera, que cada sentido diferente se corresponde con un contexto distinto. Y que los sentidos de las palabras son un subconjunto de los tipos de uso, pudiendo ser listados en un diccionario.

4.2.3 Granularidad de sentidos

Uno de los problemas más destacados en WSD es determinar el grado apropiado de la granularidad de sentidos. Algunos autores como Slator y Wilks [1987] han remarcado que la división de sentidos que uno encuentra en los diccionarios[1] es frecuentemente muy fina para las tareas propias del PLN.

Una granularidad demasiado fina dificulta la automatización del proceso WSD, al introducir unos efectos combinatorios muy significativos (por ejemplo, Slator y Wilks, muestran que la frase *There is a huge envelope of air around the surface of the earth* tiene 284 combinaciones diferentes posibles utilizando el diccionario de tamaño medio LDOCE). Esto requiere realizar una elección de sentidos extremadamente dificultosa, incluso para lexicógrafos expertos. Además, muchas veces, la distinción de significados realizada en algunos diccionarios, es difícil de hacer por parte de los lectores. Kilgarriff [1992, 1993a] muestra en un estudio, cómo a veces le es imposible al ser humano asignar muchas palabras a un solo sentido en el LDOCE. Se han propuesto distintos enfoques para reducir la cantidad de significados proporcionados por la mayoría de los lexicones. Sin embargo, no se solventa el problema, ya que la traducción automática requiere una distinción de sentidos muy fina (fina granularidad), siendo en algunos casos, más fina de la que proporcionan muchos diccionarios monolingües [Ide y Veronis, 1998]. Por ejemplo en inglés, la palabra *river*

[1] También señala el problema de la existencia de diferentes conjuntos de significados de palabras dependiendo del diccionario de que se trate.

(río) se traduce como *fleuve* en francés, cuando el río desemboca en el océano, y en otro caso como *rivière*. No hay una correspondencia estricta entre una tarea dada y el grado de granularidad requerido [Ide y Veronis, 1998]. Por ejemplo, la palabra *mouse*, aunque tiene dos significados distintos (animal, dispositivo), en ambos se traduce en francés como *souris*. Por un lado, en recuperación de información es importante la distinción entre estos dos significados de *mouse*, además es difícil imaginar una razón para distinguir *river* (sentido *fleuve*) de *river* (sentido *rivière*). Y por otro, no queda claro cuando los sentidos pueden ser combinados o expandidos, incluso para los lexicógrafos.

4.3 Investigación en desambiguación

El propósito de esta sección es proporcionar una perspectiva general del trabajo en WSD, así como hacer una breve revisión y clasificación de los métodos empleados. No se pretende exponer exhaustivamente todos los métodos, ni algoritmos existentes.

Previamente, vamos a establecer una división genérica de las metodologías seguidas en dos grandes enfoques[2]. Un primer enfoque, que podemos denominar "cualitativo", se basa en reglas para seleccionar el sentido asociado con cada palabra, así, el sistema, ante una entrada de una palabra con varias acepciones, selecciona de manera determinista el/los significado/s para los cuales las reglas se satisfacen. Por un lado, en este enfoque, las reglas específicas fallan frecuentemente en la selección ante entradas excepcionales, por otro, las reglas genéricas corren el riesgo de seleccionar sentidos incorrectos. Para contrarrestar este problema, surge un segundo enfoque, que podemos denominar "cuantitativo", el cual computa valores escalables para cada uno de los sentidos candidatos de una palabra y selecciona como sentido el de valor máximo. Este planteamiento emplea grandes bases de conocimiento y corpora de textos. Comparado con el enfoque basado en reglas, este enfoque es más robusto ante entradas excepcionales.

Por otra parte, vamos a exponer a continuación distintos enfoques de desambiguación, atendiendo al método utilizado en relación con la adquisición de conocimiento.

4.3.1 Desambiguación basada en reglas generadas manualmente

La mayoría de los sistemas de desambiguación que se desarrollaron antes de los años 80, estaban basados en reglas generadas manualmente para la selección del significado. Fue una tarea ardua la de construir este tipo de sistemas, donde el objetivo, más bien era la demostración práctica de una técnica, que propiamente un desambiguador "listo para usar" [Hirst, 1987].

Un ejemplo de esto se encuentra en [Weiss, 1973]. Weiss construyó manualmente un conjunto de reglas para desambiguar cinco palabras. Estas reglas eran de dos tipos, *reglas generales de contexto*, y *reglas de plantillas*. Una regla general de contexto debía hacer que una ocurrencia de

[2]Esta división no es radical, ni ambos enfoques son totalmente excluyentes, de hecho ha habido métodos cuantitativos con matices cualitativos. Por ejemplo, se abordaron sistemas basados en reglas sobre un conjunto de entrenamiento.

una palabra ambigua tuviese un cierto sentido, si una palabra particular aparecía cerca de la palabra ambigua. Por ejemplo, si la palabra *imprimir* aparecía cerca de la palabra *tipo*, entonces su sentido estaría relacionado con el de *impresión*. Las reglas más específicas de plantillas indicaban que una ocurrencia de una palabra ambigua tenía un cierto sentido, si una palabra particular aparecía en una localización específica relativa a esa ocurrencia. Por ejemplo, si el término *de* aparecía inmediatamente después de la palabra *tipo*, entonces el sentido de esa ocurrencia era *variedad de*.

Mediante pruebas limitadas, Weiss encontró que las reglas de plantillas, eran mejores para determinar el sentido que las reglas de contexto, así que las aplicó primero. Para crear esas reglas, Weiss examinó 20 ocurrencias de una palabra ambigua, y entonces probó manualmente las reglas creadas sobre 30 ocurrencias. Estas pruebas se realizaron con cinco palabras ambiguas obteniendo una precisión próxima al 90%.

Un desambiguador mayor fue construido por Kelly y Stone [1975], quienes crearon manualmente un conjunto de reglas para 6.000 palabras. Las reglas eran de tipo contextual, similares a las creadas por Weiss, más una serie de reglas para comprobar ciertos aspectos gramaticales de una ocurrencia de un término. En algunos casos, la categoría gramatical de una palabra es un indicador fuerte de su sentido. Por ejemplo "el pez" o "la pez". Las reglas gramaticales y de contexto fueron agrupadas en conjunto, para que sólo determinadas reglas se aplicaran en ciertas situaciones. Sentencias condicionales controlaban la aplicación de los conjuntos de reglas. A diferencia del sistema de Weiss, el desambiguador fue diseñado para procesar una frase cada vez. Podía variar el orden en el que las palabras de la frase eran desambiguadas parando el desambiguador en una palabra, intentando desambiguar otras palabras en la frase, y entonces volviendo a la palabra original para descubrir si la desambiguación podía completarse. El sistema, sin embargo, no fue un éxito y Kelly y Stone informaron:

"We applied these techniques very energetically to real human language, an it became absolutely clear that such a strategy cannot succeed on a broad scale."

Otra aproximación fue realizada por Small y Rieger [1982], usando lo que ellos denominaron "palabras expertas", que en esencia eran programas. Su idea fue construir un "experto" para cada palabra ambigua. Para desambiguar las palabras en una frase, se llamaba al experto de cada una de esas palabras. Un experto podía examinar su propio contexto, tomar decisiones acerca de los posibles significados de una palabra y hacer públicas esas decisiones a otros expertos. Si mientras procesaba sus evidencias, un experto quedaba en punto muerto, podía esperar a que otros expertos publicaran sus decisiones. Estas evidencias adicionales podían ayudar al experto en punto muerto, para finalizar la desambiguación de su palabra. No hay mención de la prueba de este desambiguador, y parece ser que el informe sobre este trabajo de Small y Rieger, no va más allá del proceso de construcción de expertos. En un punto de su trabajo ellos comentaban:

"the expert for the word 'throw' is currently six pages long . . . this is large, but it should be ten times that size"

Los desambiguadores descritos se basaban en reglas generadas manualmente para determinar los sentidos de las palabras. Cuando intentaron extender su trabajo a vocabularios más grandes, el esfuerzo de la construcción llegó a ser demasiado grande, y sus resultados finales, poco exitosos.

Sin embargo, desde mediados de la década de los 80, la investigación en desambiguación ha abordado enfoques variados, (paralelamente) desde los enfoques basados en reglas generadas automáticamente, hasta sistemas basados en el conocimiento y en corpora de textos. A continuación pasaremos a describir estos enfoques de desambiguación.

4.3.2 Métodos basados en corpora de textos

Un estudio empírico sobre los significados o sentidos de las palabras requiere no sólo un diccionario donde se proporcione un conjunto inicial de sentidos, sino ejemplos de usos de esas palabras ("banco de ejemplos"). Un corpus es un ejemplo amplio y sustancial de una lengua o sublengua (ver Capítulo 3).

Los corpora han sido estudiados y utilizados desde la segunda mitad de este siglo por los lingüistas. Algunos de estos trabajos conciernen al estudio de los significados, al de las colocaciones en inglés, al de la frecuencia de las palabras comunes en inglés, etc.

En 1980 despierta nuevamente el interés por los corpora lingüísticos [Leech, 1991]. El desarrollo y la disponibilidad de herramientas para el análisis de corpora, así como los avances tecnológicos, permitieron la creación y almacenamiento de corpora cada vez más grandes, convirtiéndose así en una fuente de información para la desambiguación. Esto ha propiciado el desarrollo de nuevos modelos, utilizando frecuentemente métodos estadísticos.

Entre estos modelos se han utilizado los llamados supervisados, en cuyo entrenamiento se emplea información proporcionada normalmente por un corpus desambiguado[3] (corpus anotado semánticamente). Estos métodos supervisados requieren conjuntos de entrenamiento de gran tamaño con anotaciones manuales de significados. Por otro lado, existen situaciones en las que no se dispone de estos recursos para poder realizar la desambiguación, por lo que se ha realizado de manera no supervisada. Estrictamente hablando, la desambiguación completamente no supervisada no es posible, si lo que perseguimos es una anotación o etiquetación de sentidos. La anotación de sentidos requiere alguna caracterización de los significados, y ésta puede venir dada por un lexicón, en el qué estén claramente definidos los sentidos. Sin embargo, la discriminación de sentidos se puede llevar a cabo, completamente, de una manera no supervisada: se puede discriminar entre dos conjuntos diferentes de objetos sin necesidad de la etiquetación [Daspa, 1999].

El primer estudio, a escala real, de este enfoque basado en corpus lo realizó Black [1988], con el desarrollo de un modelo utilizando árboles de decisión y haciendo uso de un corpus de 22 millones de términos. Anotó manualmente el sentido de 2.000 palabras y seleccionó 5 palabras ambiguas que al menos tuvieran tres sentidos para realizar las pruebas. Los experimentos dieron

[3]Cada aparición de la palabra ambigua en el corpus se encuentra anotada con el sentido apropiado conforme al contexto donde aparece.

unos resultados entre el 45% y 75% en la resolución del sentido, según las diferentes estrategias que utilizó en el entrenamiento.

Desde entonces, el aprendizaje supervisado ha sido empleado en varios trabajos ([Hearst, 1991; Gale et al., 1992a, 1993; Voorhees, 1993; Bruce y Janyce, 1994] entre otros). Sin embargo, el mayor obstáculo, a pesar de la disponibilidad de grandes corpora, ha sido la escasez de corpora de entrenamiento anotados semánticamente (ver Capítulo 3).

Así Hearst [1991], anotó los sentidos manualmente para entrenar a su desambiguador. Para evaluar su sistema con una palabra ambigua determinada, tuvo que desambiguar manualmente cierto número de ocurrencias de esa palabra. El desambiguador realizaba un análisis léxico y gramatical del contexto de las ocurrencias a desambiguar, para adquirir información que le ayudara a discriminar entre los sentidos de la palabra. Una vez que el sistema hubo pasado la fase del entrenamiento supervisado, lo intentó con el entrenamiento sin supervisar para intentar mejorar la eficiencia de su sistema[4]. El desambiguador intentaba resolver la ambigüedad de una ocurrencia y reunir la misma información léxica y gramatical del contexto de esa ocurrencia reunida en la fase del entrenamiento supervisado.

Realizó una evaluación sobre seis ocurrencias de palabras ambiguas, que ella misma había desambiguado manualmente para este propósito. Obteniendo una precisión de sus trabajos comprendida en el rango del 73% al 100%, si bien la desambiguación perfecta fue sólo para una palabra. Después de todos sus experimentos concluyó, que cuanto mayor fuese el conjunto de entrenamiento supervisado en número de palabras, mejores eran los resultados.

Entre las técnicas no supervisadas encontramos los trabajos de Brown et al. [1991] quienes extrajeron un modelo estadístico del corpus bilingüe Hansard, y Yarowsky [1992] que reúne en clases las palabras que concurren a través de un corpus no anotado.

Por otra parte, Schütze [1992] redujo la supervisión manual utilizando algoritmos de *clustering*. Éstos los utilizó, para dividir el conjunto de datos de entrenamiento en un cierto número de clusters. Una persona experta examina un pequeño número de ejemplos (de 10 a 20) contenidos en cada *cluster*, éstos se aplican, para determinar el sentido apropiado de cada *cluster*. Claramente, este método no constituye un aprendizaje supervisado, ya que usa datos de entrenamiento no anotados, sin embargo, el proceso de adquisición no es completamente no-supervisado. Para una entrada dada, se selecciona el *cluster* con valor máximo de similitud. En [Shütze y Pedersen, 1995] esta idea se extiende, al agrupar mediante clustering los vectores de contexto en clases (representando sentidos de palabras). Las ocurrencias se desambiguan asignándolas a sus clusters más cercanos.

Otros métodos se fundamentan en información contextual de carácter local, al considerar como contexto las palabras más inmediatas de su entorno (5 ó 10 ó la misma frase). Así, es expresada esta idea en [Yarowsky, 1995] basándose en el principio "un sentido por discurso":

[4]Emplea una metodología de bootstrapping. Se fundamenta en que dado un conjunto inicial de entrenamiento (usualmente consta de un pequeño número de ejemplos anotados) ir progresivamente enlazando los datos de entrenamiento, adquiridos iterativamente de los resultados obtenidos de desambiguaciones previas (supuestamente correctas).

same words are likely to have the same meanings if they occur in similar local contexts

Algunos autores que recogen esta idea son Bruce y Wiebe [1994b,a], quienes descomponen el modelo probabilístico que resultaría de tomar varias características contextuales locales (morfológicas, colocaciones, POS, ...) como interdependientes. Y Pedersen et al. [1997] que comparan tres algoritmos de adquisición del lenguaje basados en modelos estadísticos, usando características contextuales locales y globales.

Yarowsky [1995] basándose en el principio, —*un sentido por discurso y un sentido por colocación*—, emplea una algoritmo no-supervisado para clasificar las ocurrencias de una palabra dada, en una de sus posibles clases. El algoritmo consta de un procedimiento basado en corpus que reúne características de contexto local, para más tarde poder utilizarse en WSD.

Desambiguación basada en la traducción de un segundo corpus (corpus bilingüe)

Algunos autores propusieron la utilización de corpora bilingües para así evitar el etiquetado manual de los significados. Gale y Church [1991] y Dagan y Itai [1994] utilizaron recursos bilingües para WSD, basándose en la observación siguiente: diferentes sentidos de palabras en una lengua dada, pueden corresponderse con palabras distintas en otra lengua.

El algoritmo propuesto por Dagan —quien comentó— *two languages are better than one*, utiliza la correspondencia dada entre palabras en un diccionario bilingüe. La idea básica de Dagan puede resumirse con un ejemplo. En inglés la palabra *interest* tiene dos traducciones en alemán (entre otras): **Beteiligung** (*legal share* como en "a 50% interest in the company") y **Interesse** (*attention, concern* como en "her interest in mathematics"). Para desambiguar una ocurrencia de *interest* en inglés (el primer lenguaje en nuestro ejemplo), podemos identificar la frase donde ésta aparece y buscar un corpus como segundo lenguaje (en este caso en alemán) con ejemplos de la frase. Si la frase aparece con una sola traducción de la palabra *interest* en la segunda lengua, entonces se puede asignar el correspondiente sentido, siempre que el término *interest* sea utilizado en esta frase. Supongamos ahora que *interest* aparece en la frase *showed interest.* La traducción en alemán de *showed, zeigen*, aparecerá únicamente con *Interesse* si el índice de interés no puede mostrarse. Podemos concluir que el término *interest* en la frase *show interest* pertenece al sentido *attention, concern.* Por otra parte, la única traducción frecuentemente utilizada de la frase *acquired* and *interest* es *erwarb eine Beteiligung*, si *interest* no puede ser adquirido con el sentido *attention, concern.* Esto nos dice que un uso de *interest* como el objeto *acquire* corresponde al sentido, "legal share".

Por otra parte, Gale y Church [1991] empleando corpora bilingües, observaron que las estructuras gramaticales y de párrafos eran idénticas en los dos corpora, a causa de que ambos eran traducción directa el uno del otro. Utilizó una técnica automática de alineación de cada frase de un corpus con la traducción en el otro corpus, con un alto grado de precisión. El resultado fue un "corpus bilingüe alineado", que se podía utilizar para descubrir cómo una palabra en una determinada frase se traducía al otro lenguaje. Usando el mismo principio explotado por Dagan, la traducción de una palabra generalmente reflejaba los sentidos de aquella palabra, y por eso las

palabras alineadas del corpus podían ser automáticamente etiquetadas con información acerca del sentido correcto. Gale realizó pruebas muy limitadas sobre su desambiguador, ya que sólo probó con la palabra "bank", alcanzando una precisión próxima al 92%.

Los métodos de Gale y Dagan a pesar de su innovación estaban limitados por el número de sentidos que podían resolver, así como por la fiabilidad de las distinciones de sentidos que quedaban reflejadas en la traducción del lenguaje. En los informes de sus desambiguadores, ninguno de los autores hacen referencia a esta característica.

La aplicación de estos métodos basados en corpora bilingües está limitada, ya que se restringen a aplicaciones de traducción automática.

4.3.3 Métodos basados en el conocimiento

La resolución de la ambigüedad léxica utiliza información de un lexicón. El lexicón puede ser un diccionario electrónico o un thesaurus.

Desambiguación basada en diccionarios electrónicos

Como ya se ha comentado en el Capítulo 3, los diccionarios se han convertido en una popular fuente de conocimiento para muchas tareas del procesamiento del lenguaje natural. Una actividad desarrollada durante los años 80, fue aquella de intentar extraer automáticamente bases de conocimiento léxico y semántico de diccionarios electrónicos. Este objetivo —extracción automática de grandes bases de conocimiento— no se consiguió totalmente, sólo la construcción de la base de conocimiento léxico a escala real WordNet. Se han demostrado las dificultades para realizar la extracción automática de relaciones como la simple hiperonimia [Veronis y Ide, 1991], debido en parte a las inconsistencias de los diccionarios, además del hecho de que los diccionarios se construyeron para el uso humano y no para la explotación máquina.

A pesar de sus deficiencias, los diccionarios electrónicos proporcionan una fuente de información sobre las acepciones de las palabras, convirtiéndose en un elemento importante a tener en cuenta en la investigación en WSD. Los métodos empleados intentan evitar los problemas citados anteriormente, junto con métodos suficientemente robustos para reducir o eliminar los efectos de las inconsistencias dadas en los diccionarios.

Lesk [1986] propuso uno de los primeros enfoques basado en diccionarios, utilizando una idea simple: las definiciones de las palabras en los diccionarios son probablemente buenos indicadores para resolver los sentidos. Para ello, creó una base de conocimiento a partir de las definiciones textuales proporcionadas por el diccionario para cada significado. Este método conseguía entre un 50% y un 70% de desambiguaciones correctas (para un ejemplo de palabras ambiguas), utilizando un conjunto de sentidos similares a los que utilizan los diccionarios. Este enfoque es muy dependiente de las definiciones, ya que la presencia o ausencia de una palabra dada, puede alterar los resultados. Lesk reconoció este problema y mencionó que su desambiguador era incapaz de procesar varias palabras ambiguas, a causa de la no-existencia de palabras que concurrieran entre las definiciones del contexto y las palabras ambiguas. Sugiriendo que una solución podría ser,

usar diccionarios con definiciones más largas, como el "Oxford English Dictionary", sin embargo, esta idea nunca se llevó a cabo.

La importancia del método de Lesk ha significado, por un lado, la base para el trabajo en desambiguación basada en diccionarios, y por otro, ha servido para demostrar que el uso de un diccionario incrementaba la capacidad de un desambiguador, en cuanto a la resolución del significado en una gran cantidad de palabras.

Después, Wilks et al. [1990] intentaron mejorar el conocimiento asociado con cada sentido, usando una técnica de expansión de definiciones de diccionarios[5], calculando la frecuencia de concurrencia para las palabras que formaban parte de la definición. Esto lo realizó mediante un método vectorial, que relacionaba cada palabra con su contexto. En sus experimentos conseguía un 45% de precisión en la identificación del significado de una palabra simple (*bank*) que aparecía en unas 200 frases[6], con un nivel de granularidad fina (13 sentidos). Por otro lado, utilizando una granularidad gruesa (5 sentidos) obtuvo el 85% de aciertos.

El método de Lesk se extendió posteriormente en [Veronis y Ide, 1990], mediante la creación de una red neuronal a partir de las definiciones de los textos del *Collins English Dictionary* (CED), donde cada palabra se unía a sus sentidos. Los experimentos que realizaron sobre 23 palabras ambiguas, dieron una precisión próxima al 72% de los casos, utilizando la granularidad de sentidos proporcionados por el CED. En experimentos posteriores, modificando varios parámetros, mejoraron la precisión al 85% [Veronis y Ide, 1995]. Reprodujeron este método con textos completos obteniendo resultados similares (72% de asignaciones correctas comparada con el 33% del línea base y el 40% del método de Lesk).

Varios autores (Krovetz y Croft [1989]; Guthrie et al. [1991]; Slator [1992]; Cowie et al. [1992]; Janssen [1992]; Braden-Harder [1993]; Liddy y Paik [1993]) intentaron mejorar los resultados, utilizando campos de información suplementaria de la versión electrónica del LDOCE, en particular, los campos **box code** y **subject codes** para cada sentido. El campo **box code** incluye primitivas tales como ABSTRACT, ANIMATE, HUMAN, etc., y codifican tipos de restricciones de nombres y adjetivos y sobre los argumentos de los verbos. El campo **subject codes** utiliza otro conjunto de primitivas para clasificar los sentidos de las palabras por tema (ECONOMICS, AUTOMOTIVE, ENGINEERING, etc.). Así, Guthrie et al. [1991] demostró un típico uso de esta información, explotando un conjunto de categorías temáticas asignadas a muchas de las definiciones de la versión electrónica del LDOCE. El ámbito de las categorías cubre un amplio rango desde lo más general a lo más particular.

El método desambiguación empleado 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 ca-

[5]La idea era que las palabras que concurrían normalmente estaban relacionadas semánticamente con aquellas de su definición. Utilizó para expandir el LDOCE. Este diccionario se diseñó para las personas que usan el inglés como su segunda lengua, y por tanto, todas las definiciones se escribieron mediante un vocabulario simplificado de 2.000 palabras. Ya que el uso de este vocabulario produce un número mayor de concurrencias.

[6]Las frases se desambiguaron a mano con lo que el proceso de evaluación fue no sistemático y reducido.

tegoría. No están disponibles los resultados cuantitativos de la evaluación, pero Cowie et al. [1992] conjuntamente con Guthrie mejoraron el método e informaron de los resultados (47% en la distinción de sentidos y el 72% para palabras homógrafas).

Las inconsistencias en los diccionarios no son las principales limitaciones para WSD, así el conjunto de sentidos finito y discreto aportado por los diccionarios son inadecuados para resolver todas las ambigüedades como apuntan Boguraev y Pustejovsky [1990]; Kilgarriff [1991]. Asimismo, los diccionarios proporcionan información detallada en el ámbito léxico, careciendo de información pragmática que ayude a la determinación del sentido. Sin embargo, el diccionario sigue siendo una fuente de conocimiento valiosa y conveniente para la distinción de sentidos.

Desambiguación basada en thesaurus

Los thesaurus proporcionan información sobre las relaciones entre las palabras, siendo la más común la de sinonimia. El thesaurus Roget fue puesto en formato electrónico y, desde la década de los 50 se ha utilizado en una amplia variedad de aplicaciones, tales como traducción automática, recuperación de información y análisis del contenido[7]. Cada ocurrencia de una misma palabra en diferentes categorías del thesaurus, representa diferentes significados de la palabra; por ejemplo, las categorías se corresponden aproximadamente con los sentidos de las palabras, ya que las palabras pertenecientes a una misma categoría están semánticamente relacionadas [Yarowsky, 1992]. 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 tanto determina el sentido de las palabras [Manning y Schütze, 1999].

Uno de los primeros intentos lo realizaron Walker y Amsler [1986], utilizando la idea de que cada sentido y también cada palabra se asigna a una o más categorías o temas en el diccionario. Por ejemplo, para desambiguar una palabra, extraían del diccionario sus categorías, luego computaban la frecuencia de aparición de las palabras de la frase en las categorías, seleccionando el sentido para aquella categoría más frecuente.

Yarowsky [1992] construyó un desambiguador (uno de los más precisos), usando el thesaurus Roget y la Enciclopedia Multimedia Grolier. El desambiguador está basado en las categorías semánticas donde todas las palabras de Roget están emplazadas. Utilizó una amplia distinción de categorías que cubren distintas áreas (como, *herramientas/maquinaria* o *animales/insectos*), apuntando que esta información temática de bajo nivel puede proporcionar una rica información para WSD. El desambiguador de Yarowsky intentaba resolver la ambigüedad existente en las palabras de estas categorías.

La adquisición de conocimiento para decidir las categorías semánticas de una palabra ambigua, se obtenía de un etiquetado gramatical de la enciclopedia Grolier.

Para todo contexto de cada palabra se calculaba su frecuencia de aparición en la enciclopedia, asignando una puntuación basada en la comparación de frecuencias. Los experimentos que

[7] Los thesaurus, al igual que los diccionarios electrónicos, son recursos creados para el uso humano, por consiguiente no son una fuente de información totalmente coherente en cuanto a las relaciones de las palabras.

realizó, podían resolver los sentidos en un 90% de los casos para un ejemplo de 12 palabras ambiguas.

Los thesaurus proporcionan una rica red de asociaciones de palabras y un conjunto de categorías semánticas interesantes para el PLN.

Como ya se ha comentado, en la segunda mitad de los años 80, comienzan a construirse a mano, bases de conocimiento a escala real, por ejemplo, uno de los más conocidos y disponibles actualmente es WordNet[8] [Miller, 1990, 1995; Fellbaum, 1998], CyC [Lenat, 1995] y Acquilex [Verdejo, 1994].

WordNet se diseñó para utilizarse solamente en trabajos basados en computadora, por lo que, no tiene los problemas asociados con algunos de los diccionarios electrónicos mencionados anteriormente. Es uno de los recursos más utilizados en WSD en inglés. WordNet combina las características de otros recursos explotados comúnmente en el trabajo de desambiguación: incluye definiciones de términos para sentidos individuales como en un diccionario; esto define los *synsets* conjunto de sinónimos representando un concepto léxico como en un thesaurus, incluyendo también otros tipos de relaciones léxicas y semánticas. Actualmente proporciona el más amplio conjunto de información léxica en un único recurso. Otra razón posiblemente más convincente es que es el primer recurso léxico de amplia cobertura y el más utilizado, debido fundamentalmente a su libre distribución y amplia disponibilidad. Como resultado, la división de sentidos y las relaciones de WordNet probablemente impacten en el campo del PLN y más concretamente en WSD por varios años [Ide y Veronis, 1998].

Uno de los primeros intentos de explotación de parte de la información léxica existente en WordNet lo realizó Voorhees [1993] en el campo de la recuperación de información. En su aproximación, intenta disminuir los problemas que plantean, para IR, la polisemia y la sinonimia. Para ello utiliza principalmente la información existente en WordNet acerca de la relación *is-a* (hiponimia) para los nombres del inglés. Definió el *hood* de un significado o *synset* como "categoría de significados" asociado a él, algo análogo a la representación de categorías realizadas sobre el Roget en los métodos comentados anteriormente, pero adaptado a *synsets*. Para definir el *hood* de un *synset* s, considera el conjunto de *synsets* y los enlaces de hiponimia *is-a* de WordNet como el conjunto de vértices y arcos dirigidos de un grafo. Para asignar a una palabra un *hood* con un peso en un documento (o identificar su significado concreto en el contexto), Voorhees propone un criterio basado en la frecuencia de aparición de los *hoods* en la colección de documentos y en el texto considerado. De esta forma, dada una palabra en un documento y sus *hoods* asociados, le da mayor peso a los *hoods* que son más referenciados en el texto respecto al número de referencias en la colección. Este procedimiento de desambiguación, le permite asignar un mayor peso a los

[8] Aunque muchos autores clasifiquen a WordNet principalmente como un thesaurus, categorías tales como lexicón enumerativo (debido al enfoque seguido en su construcción, ya que los sentidos son explícitamente proporcionados, frente al enfoque generativo, donde se utilizan reglas de generación para derivar el preciso sentido, ya que la información semántica no está especificada) y base de datos léxica también definen lo que es WordNet. El autor ha elegido la denominación de base de datos léxica y la ha clasificado como un thesaurus, por ser esta clasificación una de las más extendidas internacionalmente. Para más información véase "Five Papers about WordNet", disponible en el área de FTP de la Universidad de Princeton.

hoods que son más referenciados en un documento, y que deben representar mejor el significado concreto del término en el documento. Una vez representados los documentos por sus *hoods* con sus pesos, se define una función de similitud análoga a la del modelo del espacio vectorial. Los resultados indicaron que la técnica no era fiable para distinguir la fina granularidad de sentidos de WordNet.

La taxonomía de WordNet puede utilizarse como fuente léxica. Así Sussna [1993] eligió la red semántica de nombres de WordNet como fuente de información para su desambiguador. La hipótesis utilizada, era seleccionar para un conjunto de términos (que aparecen cercanos unos de otros en un texto) los sentidos que minimicen las distancias entre los términos del conjunto. La red le permitió calcular la distancia semántica entre dos palabras cualesquiera de la red, y así calcular la distancia semántica para los términos (nombres) a desambiguar. Para realizar esto, él asignó pesos a varias relaciones (sinonimia, hiperonimia, hiponimia, etc.) de la red semántica. El valor del peso asignado a la relación reflejaba la similitud semántica expresada por esa relación. Por ejemplo, a la relación de sinonimia asociada a un *synset* se le asignó los pesos mayores y a la de antonimia los pesos menores. La distancia semántica entre dos nodos se calculaba sumando los pesos unidos de las relaciones que hicieran el camino más corto entre los dos nodos y se seleccionaban los sentidos que minimizaban la distancia. Sussna no hizo mención del precio potencial que suponía buscar este camino mínimo en la red semántica y los resultados obtenidos de una evaluación manual fueron del 56%. Pero este trabajo es particularmente interesante, porque es el único que utiliza otras relaciones léxicas además de la hiponimia.

Agirre y Rigau [1996] propusieron un método para la resolución de la ambigüedad léxica de nombres. El método hacía uso de la taxonomía de nombres de WordNet y de la noción de densidad conceptual entre conceptos, una extensión de distancia conceptual[9]. Asimismo, Resnik [1995a,b] exploró la medida de similitud semántica (estrechamente relacionada con la distancia conceptual) por significados en la jerarquía de WordNet. La distancia conceptual proporciona una base para determinar la cercanía en relación con el significado de dos palabras, tomando como referencia una red jerárquica estructurada. La distancia conceptual entre dos conceptos se definía como el camino más corto que conecta los conceptos en la red. El método era totalmente automático y su evaluación se realizó sobre SemCor, obteniendo unos resultados prometedores (71%), al ser superiores a los obtenidos por los métodos de Sussna y Yarowsky.

Por otra parte, Rada y Moldovan [Ide y Veronis, 1998] basándose en la misma idea, distancia conceptual (o densidad semántica entre palabras), presentaron un método de desambiguación para nombres y verbos en el contexto de WordNet, obteniendo unos resultados[10] según sus estimaciones entre un 70 y un 76%.

WordNet no es el recurso ideal, ni perfecto para WSD. El problema frecuentemente más citado es la fina granularidad de sus sentidos, que va frecuentemente más allá de lo que puedan

[9]La noción de distancia conceptual entre nodos de la red, fue empleada por [Sussna, 1993], como se ha visto en el párrafo anterior.

[10]El método según comentaron era difícilmente comparable con otros, porque consideraban el significado colectivo de dos o más palabras.

necesitar muchas aplicaciones del PLN. Sin embargo, no está claro el nivel deseado en la distinción de sentidos para WSD, o incluso si este nivel es representado en la jerarquía de WordNet, o si debe ser el mismo para todas las categorías de palabras o para todas las aplicaciones del PLN. Actualmente, en la comunidad del PLN se están estudiando estos temas, incluyendo el concepto de "sentido" (ver Introducción).

Desambiguación basada en la combinación de fuentes conocimiento (híbridos)

Hay una nueva tendencia en desambiguación, que es la de combinar varias fuentes de conocimiento como un lexicón, heurísticas, colocaciones y otras.

Wilks y Stevenson [1997] proponen un sistema de desambiguación que utiliza varios *taggers* parciales, usando cada uno fuentes 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. Han implementado un *tagger* que incorpora como fuentes de información: *Part-Of-Speech*, definiciones de diccionarios y códigos de dominio (categorías de thesaurus). Un proceso final, utilizando un mecanismo muy simple, combina los resultados de los procesos que utilizan cada una de las mencionadas fuentes de información. Ellos comentaron que desambiguaban correctamente el 88% de las palabras polisémicas.

Por otra parte, Ng y Lee [1996] presentaron un enfoque haciendo uso de un algoritmo basado en el aprendizaje. Este enfoque integra un conjunto de fuentes de conocimiento para desambiguar el significado, incluyendo *Part-Of-Speech*, formas morfológicas y palabras concurrentes, colocaciones locales y relaciones sintácticas. Rigau et al. [1997] presentaron un método que puede emplearse para desambiguar en un corpus completamente no-etiquetado. Combinaron tanto un conjunto de algoritmos no supervisados como varias heurísticas (8 en total), muchas de ellas estadísticas, utilizando recursos léxicos, tales como WordNet.

Otros métodos que pueden considerarse también híbridos, son aquellos que combinan más o menos recursos léxicos con algoritmos de aprendizaje máquina. Ejemplos de éstos son los trabajos de: Siegel [1997] quién utilizó algoritmos de aprendizaje máquina para clasificar verbos, y el de Mooney [1996] que comparó siete algoritmos de aprendizaje clásicos (incluyendo redes neuronales, técnicas estadísticas y árboles de decisión) en la tarea de desambiguar entres seis sentidos la palabra *line*, utilizando información local.

4.4 Efectividad del proceso de desambiguación

Una cuestión evidente ante la existencia de distintos y variados enfoques WSD, sería decidir cual de ellos es mejor, de manera que pudiésemos realizar una comparación entre ellos. La evaluación de los sistemas WSD ha supuesto un problema en la investigación de la desambiguación, de hecho, algunos desambiguadores han sido evaluados utilizando métricas pocas extendidas y sólo a través de pruebas manuales sobre un grupo reducido de palabras, con lo que podemos afirmar que la

evaluación ha sido muy heterogénea. Así, Gale, Church, y Yarowsky [1992b] presentan una extensa discusión del problema de la evaluación de los sistemas WSD, revisando los primeros trabajos realizados, y destacando que algunas palabras son difíciles para algunos programas WSD, otras fáciles, valorando la efectividad del programa a través de un ejemplo aleatorio. Ellos consideran que hay que realizar el proceso de evaluación con precaución:

"... There are many potentially important differences including different corpora, different words, different judges, differences in treatment of precision, and differences in the use of tools such as parsers and part of speech taggers, etc."

Unos años más tarde, Resnik y de nuevo Yarowsky [1997], disertan sobre la evaluación apuntando 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. Asimismo, afirman que, la desambiguación presenta diferentes relaciones entre las distintas tareas (por ejemplo, la recuperación de información puede funcionar mejor con un enfoque WSD bastante diferente, que el que requiere la traducción automática). Concluyen, proponiendo el establecimiento entre la comunidad de un conjunto de datos de evaluación o corpus estándar ("Gold Standard Datasets" [Kilgarriff, 1998]).

En cuanto a las métricas empleadas, se han venido utilizando las existentes en el campo de la recuperación de información, con una interpretación adecuada. Así, en la bibliografía [Rijsbergen, 1979; Salton y McGill, 1983; Bollmann, 1983; Raghavan et al., 1989; Salton, 1991b] se puede encontrar un gran número de aspectos y parámetros utilizados para caracterizar el comportamiento de estos sistemas. Disponer de criterios para la evaluación de un sistema o proceso, es una cuestión determinante a la hora de enjuiciar un determinado modelo o sistema concreto, o a la hora de comparar varias propuestas.

En la evaluación de un sistema se pueden diferenciar aspectos relacionados con la efectividad y aspectos relacionados con la eficiencia [Salton y McGill, 1983; Salton, 1989]. 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, del espacio en disco ocupado por archivos auxiliares, etc. Nuestro estudio se encuentra orientado principalmente a aspectos relacionados con la efectividad. La efectividad de un sistema se centra en la precisión con que éste resuelve, en el caso que nos ocupa, la ambigüedad léxica.

Los dos índices relacionados con la efectividad que más ampliamente se han utilizado en IR son los denominados *precision* y *recall*[11]. En el caso de los sistemas de recuperación de información, el *recall* favorece a los sistemas que recuperan muchos documentos como posibles a una consulta realizada (desconsiderando el ruido contenido en los datos recuperados). Mientras que la *precision* favorece a los sistemas que recuperan pocos documentos irrelevantes como posibles. Como se puede observar cuando todos los documentos se recuperan el *recall* asciende al 100%,

[11] Utilizaremos estos términos en inglés dada la universalidad de su aceptación.

potencialmente sacrificando la *precision*. Formalmente se definen como sigue en las fórmulas 4.1 y 4.2:

$$precision = \frac{documentos\ relevantes\ recuperados}{documentos\ recuperados} \tag{4.1}$$

$$recall = \frac{documentos\ relevantes\ recuperados}{documentos\ relevantes\ en\ la\ coleccion} \tag{4.2}$$

Por otro lado, en el caso de un sistema de categorización de texto, el *recall* favorece a los sistemas que asignan a los documentos muchas categorías correctas como posibles, mientras que la *precision* favorece a aquellos que asignan pocas categorías incorrectas a los documentos como posibles[12].

En la resolución de la ambigüedad léxica hemos utilizado la precisión como métrica básica para computar la efectividad de nuestros experimentos, ya que los parámetros *precision* y *recall* pueden adaptarse a las distintas operaciones del análisis del contenido [Lehnert y Sundheim, 1991; Passonneau y Litman, 1993]. La *precision* se define como el ratio entre las palabras desambiguadas correctamente y el total de desambiguaciones:

$$precision = \frac{palabras\ desambiguadas\ correctamente}{palabras\ desambiguadas} \tag{4.3}$$

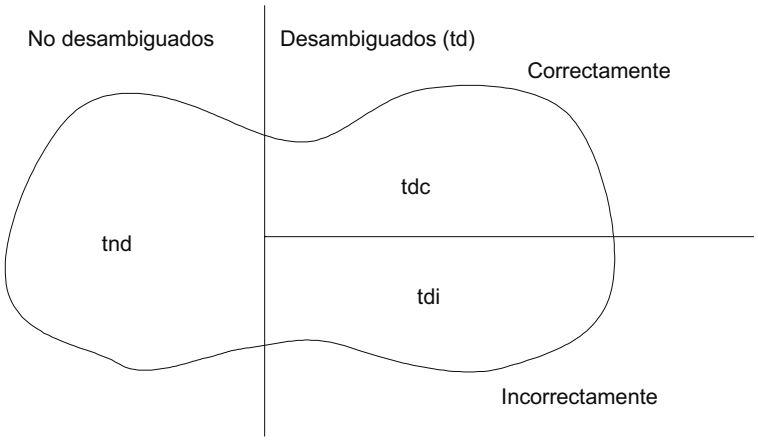

Figura 4.1: Partición de términos

[12]He de hacer notar, dependiendo del punto de vista, que un documento puede ser asignado simultáneamente a varias categorías.

En términos de la interpretación de la figura 4.1:

$$precision = \frac{tdc}{tdc + tdi} \tag{4.4}$$

El *recall* se define como el ratio entre las palabras desambiguadas y el total de palabras ambiguas que existen en la colección:

$$recall = \frac{palabras\ desambiguadas}{palabras\ ambiguas} \tag{4.5}$$

que podemos también escribir, en términos de la figura 4.1:

$$recall = \frac{td}{tnd + tdc + tdi} \tag{4.6}$$

Puede verse que, por definición, los valores de los dos parámetros se encuentran entre 0 y 1. El valor del *recall* cuantifica la capacidad del sistema de desambiguar el mayor número de posible términos. Un valor de *recall* bajo, supone que el sistema ha desambiguado una proporción reducida de términos. Algunos métodos de desambiguación son capaces de decidirse siempre por algún significado, para estos métodos el *recall* es siempre igual a 1, lo que supone que el sistema ha desambiguado todos los términos y sólo tiene sentido utilizar la *precision* como métrica de evaluación. En estos casos se calcula habitualmente la media de la *precision* para cada palabra (*macroaveraging*) y la *precision* media sobre todas las palabras (*microaveraging*) [Lewis, 1992]. El valor de *precision* cuantifica la capacidad del sistema de no presentar desambiguaciones incorrectas. Un valor de la *precision* bajo supone que el sistema ha desambiguado un gran número de palabras incorrectamente. Un valor de *precision* = 1, supone que todas las desambiguaciones realizadas por el sistema son correctas.

Los métodos de desambiguación presentados en este trabajo casi siempre toman una decisión sobre el significado a seleccionar, por tanto, el *recall* va a ser 1, como ya se ha comentado, por lo que el cálculo se va a realizar utilizando *macroaveraging* y *microaveraging*. La *precision* puede ser definida como el cociente entre el número de palabras o términos desambiguados satisfactoriamente y el número de términos desambiguados. El *macroaveraging* consiste en calcular la *precision* para cada uno de los términos, y luego calcular la media para cada uno de ellos (ver fórmula 4.7); y el *microaveraging* en calcular un solo valor de *precision* medio para todos los términos (ver fórmula 4.8).

$$P_{macroavg} = \frac{\sum P_i}{n}; \qquad P_i = \frac{tdc_i}{tdc_i + tdi_i} \tag{4.7}$$

$$P_i = Precision\ del\ término\ i$$

Donde tdc_i es el número de desambiguaciones correctas del término i, tdi_i el número de desambiguaciones incorrectas del término i y n el número de términos desambiguados.

$$P_{microavg} = \frac{tdc}{tdc + tdi} \tag{4.8}$$

Donde *tdc* es el número de términos desambiguados correctamente y *tdi* el número de términos desambiguados incorrectamente.

Como hemos comentado los índices *precision* y *recall* han sido los más utilizados para el estudio de la efectividad, pero también han sido los más sometidos a crítica. La suposición que se hace en la definición de ambos parámetros en IR de que se pueda asignar la relevancia o irrelevancia de los documentos a una consulta de una forma "objetiva", exterior al sistema ha sido frecuentemente cuestionada [Salton, 1991b]. También se ha criticado la utilización de *recall* y *precision* porque para determinados propósitos es más difícil el estudio mediante dos parámetros simultáneamente que mediante sólo uno. En este sentido, se han propuesto medidas alternativas tales como el *normalized recall, fallout, generality, sliding ratio, expected search length* y *E-measure*, entre otras [Rijsbergen, 1979; Salton y McGill, 1983; Bollmann, 1983; Raghavan et al., 1989; Salton, 1991b; Hull, 1993].

Sin embargo, pese a sus posibles deficiencias, *recall* y *precision* proporcionan información significativa y de directa interpretación, y constituyen los dos índices más utilizados para el estudio de la efectividad de los sistemas en la actualidad [Hull, 1993].

4.4.1 Evaluación de la desambiguación

Como se ha comentado, se ha producido un importante esfuerzo en la comunidad lingüística en la estandarización del proceso de evaluación de la desambiguación [Resnik y Yarowsky, 1997]. Todo este esfuerzo se centra en una evaluación de la desambiguación, que podemos denominar *directa*, es decir, en evaluar la efectividad en la asignación de los significados correctos a las palabras a desambiguar. Sin embargo, la desambiguación sirve fundamentalmente como ayuda a otras tareas, por tanto, es deseable realizar también una evaluación *indirecta*, que mida la efectividad de la tarea a la que se aplica, dependiendo del método de desambiguación empleado [Wilks, 1998].

Nosotros consideramos igualmente importante la evaluación indirecta de esta tarea [Gómez et al., 1999] como se muestra en el Capítulo 5.

Evaluación directa

La evaluación directa mide la efectividad en la asignación de los significados correctos a las palabras a desambiguar. La evaluación directa es fundamental para cuantificar la calidad de los distintos enfoques de desambiguación. Las evaluaciones directas realizadas suelen presentar los siguientes problemas:

- La falta de acuerdo en la elección de las definiciones de las palabras: diferentes diccionarios suelen proporcionar distintos conjuntos de sentidos para la misma palabra.

- La escasez de colecciones de evaluación: el etiquetado semántico de un corpus es una tarea difícil y costosa.

- La inconsistencia en el etiquetado de las colecciones de evaluación: distintas personas pueden asignar diferentes significados a la misma palabra en el mismo contexto.

- Una cierta falta de acuerdo en las métricas utilizadas: diversos autores presentan diferentes formas de medir la efectividad de la desambiguación.

Dentro de la evaluación directa podemos distinguir dos tipos de evaluación: real y artificial, que pasamos a continuación a describir.

Evaluación real: SENSEVAL[13] ha resuelto parcialmente estos problemas. En primer lugar, se ha definido una colección de evaluación única para la "competición", y paralelamente un conjunto de significados adaptados a la colección. El corpus HECTOR[14][Atkins, 1993] ha sido adaptado a SENSEVAL reanotando sus documentos con los nuevos significados definidos a partir del propio corpus. Por otra parte, la construcción de nuevos significados ha mejorado la consistencia en el etiquetado de la colección [Kilgarriff, 1998]. Por último, las métricas utilizadas han sido *recall* y *precision*.

Desgraciadamente, la evaluación realizada en SENSEVAL es demasiado genérica, independiente de la tarea a la que se aplica la desambiguación. Cada tarea posee su propia idiosincrasia, lo que impide utilizar los recursos aportados por SENSEVAL si no son adecuados a la tarea. Este es el caso de nuestro trabajo de desambiguación para la categorización y recuperación de información.

Evaluación artificial: Si no disponemos de una colección desambiguada para poder realizar la evaluación del desambiguador tenemos la tediosa tarea de desambiguar manualmente las palabras de la colección de prueba. Existe un recurso como son las *pseudopalabras*[15] (pseudo-words) [Yarowsky, 1993; Sanderson, 1996] que hacen más fácil tanto la creación de un entrenamiento a escala real, como el conjunto de evaluación o prueba para la desambiguación, obviando el anotado manual comentado anteriormente, ya que introducen ambigüedad en la colección.

[13] SENSEVAL es una "competición científica" sobre desambiguación [Kilgarriff, 1998] que se ha celebrado recientemente al estilo de la agencia norteamericana ARPA. Esta competición propone una colección de evaluación, un conjunto de definiciones y una serie de métricas concretas, y logra una alta consistencia en el etiquetado de la colección.

[14] HECTOR ha sido desarrollado conjuntamente por Oxford University Press/Digital project.

[15] Palabras ambiguas creadas artificialmente por la unión de dos o más palabras. La creación de una pseudopalabra de tamaño 2, se realiza reemplazando todas las ocurrencias de las dos palabras, por ejemplo *'bank'* y *'orange'*, por una nueva palabra *'bank/orange'*. (muestra de tamaño 3 *'bank/orange/person'*). El propósito con que creó Yarowsky las *pseudopalabras* fue la de evaluar desambiguadores.

Evaluación indirecta

La evaluación indirecta de la desambiguación mide la efectividad de la tarea 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. La evaluación indirecta es fundamental para cuantificar la calidad de los distintos enfoques de desambiguación sobre la tarea a la que se aplica.

La desambiguación se ha aplicado en otros trabajos a diversas tareas de clasificación de texto, donde los criterios de evaluación (colecciones, métricas) son los propios de cada tarea.

En el capítulo siguiente, describiremos el procedimiento de aplicación de la desambiguación a dos tareas concretas de clasificación automática de documentos, como son, la categorización de documentos y la recuperación de información.

4.5 Sistema WSD

El mecanismo básico de nuestro sistema desambiguador está fundamentado en el modelo del espacio vectorial. El modelo del espacio vectorial ha sido utilizado ampliamente en muchos trabajos de recuperación de información [Lewis, 1992; Salton y McGill, 1983; Salton, 1989], así como en otros de categorización de textos [Buenaga et al., 1997]. Basados en estas experiencias, presentamos una adaptación del modelo del espacio vectorial a la desambiguación de textos y los caminos seguidos para calcular algunos elementos del modelo.

4.5.1 Idea básica

La hipótesis básica que utilizamos es considerar que cada sentido[16] diferente se corresponde con un contexto distinto. De acuerdo con lo anterior, el sistema hace uso de la información contenida en los textos (información contextual[17]) para computar el grado de pertenencia del término a cada significado. La hipótesis clave cuando utilizamos una colección de entrenamiento para la resolución de la ambigüedad léxica es que un término aparece con un particular sentido en un determinado contexto. Así, un conjunto de términos manualmente etiquetados con el sentido correcto, puede ser utilizado para predecir el significado de nuevos términos, de acuerdo con el aforismo atribuido a Firth [1957][18]. Los términos que constituyen ese contexto pueden ser buenos para predecir el sentido con que aparece el término.

El conjunto de términos que pueden predecir el significado de un término, y su importancia, es computado estadísticamente por las ventanas contextuales [Ureña et al., 1997, 1998b], como un paso inicial del proceso de entrenamiento. Para ello, se representa cada término del corpus

[16] Como se ha visto en la sección 2, un "uso" se utiliza para designar a una ocurrencia o aparición de una palabra en un determinado contexto [Kilgarriff, 1993b]. Si tenemos dos frases que contienen una palabra o incluso en la misma frase en dos lugares diferentes, podemos decir que tenemos dos usos.

[17] El contexto semántico se ha venido utilizando en el estudio del procesamiento léxico, en el reconocimiento de una determinada palabra y en el acceso léxico.

[18] Ver Capítulo 3.

de entrenamiento con un vector, cuyas componentes son: el peso del término en el párrafo y los pesos de los términos que constituyen la ventana contextual. Así, para cada uno de los nombres de la colección de entrenamiento, calcularemos su ventana contextual, construyendo tantas como palabras con diferentes sentidos existan en la colección.

Elaboración de ventanas contextuales

El método de elaboración de ventanas contextuales varía según el corpus de textos de que se trate. Así pues, en los corpora anotados morfosintácticamente es posible elegir como unidades contextuales el párrafo[19], mientras que en corpora no-anotados, que carezcan de este tipo de etiquetado, es necesaria su construcción basándonos en unidades contextuales de tamaño frase (si la información de separación de frases está disponible) o simplemente en conjuntos de términos (por ejemplo entornos de términos).

La forma de construcción de las ventanas contextuales influye en los resultados obtenidos, puesto que de ella depende el uso que hagamos de la información subyacente en el contexto del término para el que construimos su ventana contextual. La forma más natural de elaboración de ventanas contextuales para términos, suele ser coger todas las palabras no vacías pertenecientes al mismo párrafo al que pertenece el término en cuestión. De esta forma, consideramos que el contexto de una palabra es su párrafo, hipótesis lógica, puesto que el párrafo es considerado como una unidad semántica dentro de un documento.

Todas las ventanas contextuales obtenidas como consecuencia del procesamiento de la colección de entrenamiento, se agrupan en lo que denominamos *tabla de entrenamiento*, en la que se encuentran ordenadas según la palabra a la cual pertenecen.

Hay que destacar que la construcción de las ventanas contextuales puede variar según el corpus utilizado en el entrenamiento, ya que depende del formato del corpus utilizado, podemos tener corpora anotados y no anotados.

El proceso de construcción de ventanas contextuales puede variar según el corpus utilizado en el entrenamiento. A pesar de que en esta memoria sólo se indica la utilización del corpus SemCor, debido a su anotación semántica y a su posterior evaluación automática, así como a su integración con WordNet, el proceso de construcción de ventanas contextuales se ha generalizado a cualquier corpus de textos, teniendo en cuenta tanto corpora anotados como no-anotados[20].

4.5.2 Metodología

El modelo WSD que presentamos está situado en gran medida dentro modelo del espacio vectorial para la recuperación de la información. El modelo del espacio vectorial constituye el más usado hasta la actualidad para la realización de experimentos en IR [Salton y McGill, 1983; Salton,

[19] En este caso las ventanas contextuales se corresponderían con los párrafos del documento.

[20] En corpora no-anotados y sin información referente al párrafo, hemos utilizado como ventanas contextuales un entorno de n frases alrededor del término considerado.

1989]. Por otra parte, nuestro modelo WSD es un modelo integrador, puesto que facilita la utilización de diversos recursos lingüísticos en la desambiguación. El objetivo final del modelo es permitir la integración de recursos para mejorar la efectividad de la desambiguación.

Elementos fundamentales del modelo

Nosotros utilizamos el modelo del espacio vectorial para representar el lenguaje natural por medio de vectores de pesos. Cada peso representa la importancia de un término, en relación con un determinado sentido en la expresión del lenguaje natural. Cada término $\vec{s}_{ji}$ queda representado o indexado por un vector de dimensión m, con los pesos asignados a cada uno de los términos de indexación. El término i con sentido j, queda representado con el peso del término, así como con los pesos de los términos circundantes, esto es lo que denominamos *ventana contextual*. Con este concepto, hacemos referencia a las palabras que circundan al término a desambiguar, es decir, a las palabras que están en su contexto, ya que pueden suministrar información acerca del sentido utilizado.

$$\vec{s}_{ji} = \langle ws_{j1}, ws_{k1}, ..., ws_{kn} \rangle \tag{4.9}$$

ws_{kc} peso de la palabra circundante c al término $\vec{s}_{ji}$.

Para el procesamiento de los textos a desambiguar, se obtienen los términos de indexación aparecidos en ellos, de una forma análoga al de lo textos de la colección de entrenamiento. La representación de una consulta de un término $\vec{c}_k$, se realiza mediante un vector de pesos asociados a los términos.

$$\vec{c}_k = \langle ws_{c1}, ws_{ck1}, ..., ws_{ckn} \rangle \tag{4.10}$$

wc_{kc} peso de la palabra circundante c al término $\vec{c}_k$.

De esta forma, se propone la definición de una *función de similitud* (similarity) entre el término i con sentido j y el término a desambiguar. De acuerdo con las palabras de Tversky [1977]:

"Similarity plays a fundamental role in theories of knowledge and behavior. It serves as an organizing principle by which individuals classify objects, form concepts, and make generalizations"

En la figura 4.2 aparece una representación gráfica de vectores representantes de sentidos y consultas[21]. Nótese que la dimensión del espacio vectorial es, en general, m. El modelo del espacio vectorial conlleva la suposición básica, de que la similitud semántica entre los objetos representados viene dada por el coseno del ángulo que forman sus vectores.

[21] Las consultas representan los términos a desambiguar.

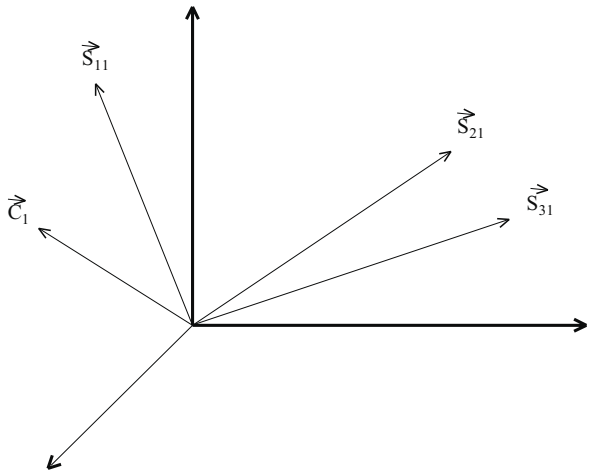

Figura 4.2: Vectores representando significados y consultas

Para desambiguar un término $\vec{c}_k$, calculamos la similitud entre el contexto en el que aparece y las definiciones de los términos. Se selecciona el sentido que presenta mayor similitud con el contexto, con arreglo a la fórmula 4.11:

$$sim(\vec{s}_{ji}, \vec{c}_i) = \frac{\sum_{i=1}^{m} ws_{ji} \cdot wc_i}{\sqrt{\sum_{i=1}^{m} ws_{ji}^2 \cdot \sum_{i=1}^{m} wc_i^2}} \tag{4.11}$$

El proceso de desambiguación es similar a la recuperación por *ranking*: dada una ocurrencia de una palabra ambigua, las definiciones de los sentidos de la palabra eran tratados como pequeñas colecciones de documentos y las palabras del contexto eran tratadas como una consulta. A cada definición le era asignada una puntuación basada en el número de definiciones de palabras encontradas en el contexto. Las definiciones eran agrupadas según sus puntuaciones y las mejor puntuadas eran elegidas para determinar el sentido correcto.

4.5.3 Algoritmos de aprendizaje

Hemos elegido los algoritmos de Rocchio [Rocchio, 1971] y Widrow-Hoff [Widrow y Sterns, 1985] para computar los pesos de los términos para un determinado significado en nuestro enfoque, como se muestra a continuación. El primero es un algoritmo empleado tradicionalmente para realimentación por relevancia en recuperación de información. El segundo es un algoritmo de aprendizaje máquina. Ambos dan la oportunidad de integración a través de una representación inicial computada por la utilización de un recurso externo como WordNet [Buenaga et al., 1997].

Mostramos, cómo calcular los vectores de pesos para cada término utilizando los dos algoritmos. Suponemos la existencia de un conjunto P de documentos de entrenamiento, previamente representados utilizando la fórmula 4.14.

Algoritmo de Rocchio

El algoritmo de Rocchio produce un nuevo vector de pesos wc_k de uno existente wc_k^0 y una colección de documentos de entrenamiento. El componente i del vector wc_k se calcula por la fórmula:

$$wc_{ik} = \alpha wc_{ik}^0 + \beta \frac{\sum_{l \in C_k} wd_{il}}{n_k} + \gamma \frac{\sum_{l \notin C_k} wd_{il}}{P - n_k} \tag{4.12}$$

Donde wc_{ik}^0 es el peso inicial del término i para el significado k, wd_{il} es el peso del término i para el ítem l de entrenamiento, C_k conjunto de índices de los elementos asignados al significado k, y n_k el número de estos elementos. Los parámetros α, β y γ controlan el relativo impacto de los pesos inicial, positivo y negativo respectivamente en el nuevo vector.

Como Lewis et al. [1996], hemos usado los valores $\beta = 16$ y $\gamma = 4$. El valor de α se establece a 20, para equilibrar la importancia de los pesos iniciales y de entrenamiento. Restringimos el clasificador para no hacer uso de pesos negativos, así al final el peso wc_{ik} será positivo, o retornará a 0 si es negativo.

El vector inicial wc_k^0 es tomado frecuentemente como vector nulo, pero esto puede ser establecido con un conjunto de pesos iniciales calculados por la utilización de un recurso externo. En la siguiente sección, veremos como se hace esto empleando WordNet.

Algoritmo de Widrow-Hoff

El algoritmo de Widrow-Hoff comienza con un vector de pesos existente wc_k^0 y secuencialmente se va actualizando una vez para cada ítem de entrenamiento. El componente i del vector wc_k^{l+1} es obtenido del ítem l y del vector l por la fórmula:

$$wc_{ik}^{l+1} = wc_{ik}^l + 2\eta(wd_l \cdot wc_k^l - y_l)wd_{il} \tag{4.13}$$

Donde wc_{ik}^l es el peso del término i en el vector l para la clase k, wd_l es el vector de pesos del término i para el ítem l, wc_k^l es el vector l para la clase k, y y_l es 1 si el ítem l es asignado a la clase k y 0 en otro caso, y wd_{il} es el peso del término i en el ítem l. La constante η es el ratio de aprendizaje, el cual controla cómo de rápido le está permitido cambiar el vector de pesos y cuanto influye cada nuevo ítem sobre éste. Un valor típicamente usado para η es $\frac{1}{4X^2}$, siendo X el valor máximo de los vectores que representan los elementos de entrenamiento.

Como en el algoritmo de Rocchio, un vector inicial de pesos se puede producir utilizando un recurso independiente, sin embargo, la importancia de este peso se reduce proporcionalmente al número de elementos de entrenamiento disponibles para una clase. Cuando hay muchos ejemplos de entrenamiento, el peso inicial es dominado por el peso obtenido de estos ejemplos, pero cuando hay peso inicial tiende a mantener sus valores.

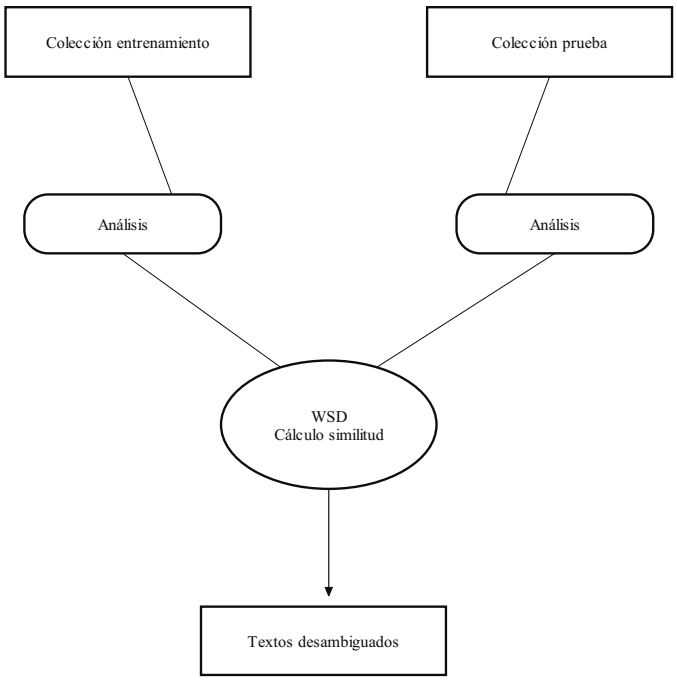

Figura 4.3: El proceso de desambiguación

4.6 Enfoques WSD

4.6.1 Desambiguador basado en corpus

La técnica utilizada para la resolución automática de la ambigüedad léxica, se enmarca dentro de la adquisición del significado de las palabras, mediante representaciones contextuales. Una representación contextual es una caracterización del contexto lingüístico, en el que una palabra expresa un determinado sentido [Miller y Charles, 1991].

El sistema hace uso de la información contenida en los documentos para determinar el significado. Como es lógico, no todos los términos tendrán el mismo número de acepciones, sino que éste será variable, dependiendo de la palabra en cuestión. Los sentidos están representados con etiquetas numéricas que codifican el sentido en WordNet.

En la figura 4.4 se ilustra la arquitectura del algoritmo anotador basado en corpus, haciendo uso del corpus SemCor.

Entrenamiento con ventana contextual

Representamos cada término del corpus de entrenamiento, por medio de un vector, cuyas componentes son: el peso del término en el párrafo; y los pesos de los términos que constituyen lo que hemos denominado la *ventana contextual*. Con este concepto, hacemos referencia a las palabras que circundan al término a desambiguar, es decir, a las palabras que están en dicho contexto,

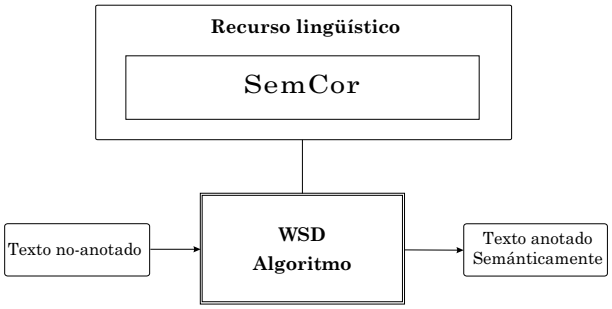

Figura 4.4: Aquitectura del enfoque WSD basado en el recurso lingüístico SemCor

ya que pueden suministrar información acerca del sentido utilizado. Así, mediante las ventanas contextuales hacemos uso obligado de la hipótesis clave apuntada anteriormente, donde:

"todas las ocurrencias de un término en un contexto comparten el mismo significado en dicho contexto"

Las *ventanas contextuales* pueden ser encuadradas dentro del Modelo del Espacio Vectorial considerándolas como los vectores que dicho modelo maneja.

Para ello es necesario indexar los documentos en ventanas contextuales, y éstas a su vez, en vectores capaces de ser tratados en un espacio vectorial. En la figura 4.5 aparece una representación gráfica de vectores representantes de ventanas contextuales.

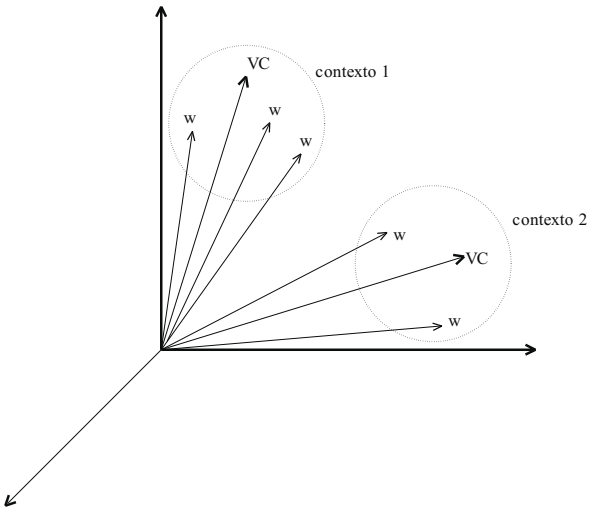

Figura 4.5: Vectores representando *ventanas contextuales*

Así, para cada uno de los nombres de la colección de entrenamiento, se calcula su ventana contextual. El programa desplaza la ventana, desde el principio de todos los documentos que

contiene el corpus de entrenamiento, hasta el final de los mismos, considerando en cada desplazamiento un nombre, y como palabras de contexto, cada una de las palabras circundantes (ver figura 4.6). De esta manera, se construyen tantas ventanas contextuales como términos con diferentes sentidos existan en la colección. El tamaño de la ventana contextual será variable y estará en función del número de términos que contenga el párrafo en cuestión.

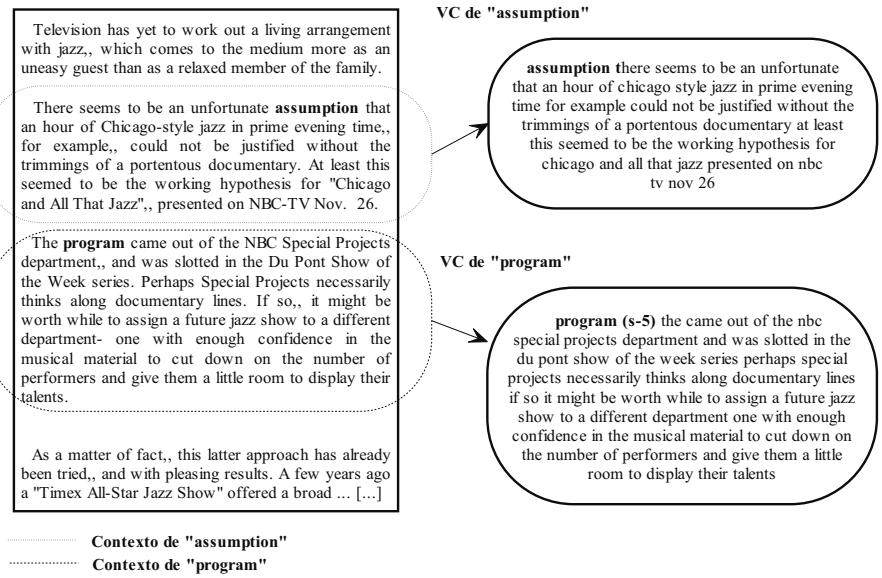

Figura 4.6: Ejemplo de Ventana Contextual (fragmento extraído del documento *br-c02* de SemCor)

Una vez realizado esto, se construyen los vectores para la colección de entrenamiento y se calculan los pesos para los distintos términos de manera análoga a Salton y McGill [1983]:

$$ws_{ji} = t_{ji} \cdot w_i \tag{4.14}$$

$$\text{siendo } w_i = \log_2(\frac{n}{f_i})$$

Donde t_{ji} es la frecuencia del término j con sentido i en la ventana contextual, n es el número de sentidos del término i, y f_i es el número de ventanas contextuales donde aparece el término i. La dimensión del espacio vectorial es variable y está en función del número de sentidos que tenga la palabra a desambiguar.

Se suman los vectores que representan el mismo término y el mismo sentido, para que cada término i con sentido j quede representado por un solo vector, dada la susceptibilidad de repetición de algunos de ellos. Esto se realiza como consecuencia de la construcción de los vectores s_{ji}, para cada uno de los nombres i con sentido j que tiene el corpus.

El principal problema de este enfoque se produce cuando no se dispone de suficiente información contextual en el entrenamiento, por lo que habrá que obtenerla mediante un recurso lingüístico externo.

4.6.2 Desambiguador basado en WordNet

En el estudio que nos ocupa, se puede utilizar una base de datos léxica como WordNet, para la identificación del significado de las palabras en un determinado contexto. Recientemente muchas de las investigaciones en WSD que se han realizado, se han basado en el entrenamiento con corpus (como se puede ver en el Capítulo 3).

Sin embargo, los datos de entrenamiento a veces son escasos en los enfoques basados en corpora. Uno de los objetivos de esta investigación es el uso de las distintas relaciones de WordNet para incrementar la efectividad de los datos de entrenamiento.

Explotando las relaciones léxicas y semánticas conceptuales de WordNet en la desambiguación

WordNet [Miller, 1995; Landes et al., 1998] proporciona un número importante de relaciones conceptuales, léxicas y semánticas, además de la *sinonimia* y *antonimia* que poseen muchos de los thesaurus existentes. Estas relaciones se pueden utilizar para ingeniar un cambio en la representación de la información textual, transformando los vectores de palabras en vectores de significados. Así por ejemplo, la relación de *sinonimia* se utiliza para clasificar las palabras con significados similares (o usos similares). Y la relación de *hiperonimia*[22] (que corresponde a la relación "es un") se utiliza para generalizar los significados de los nombres en el nivel más alto de abstracción, como se ilustra en figura 4.7.

En otros trabajos de desambiguación se analiza la expansión de términos que denominamos "expansión de sentidos" con información meramente de *sinonimia* [Ureña et al., 1998a; Agirre y Rigau, 1996] en el entrenamiento. Sin embargo, en este enfoque se estudia el aporte de las propias relaciones de WordNet a WSD, no sólo mediante información de *sinonimia*, sino a través de diferentes relaciones léxicas y semánticas como la *hiperonimia, meronimia, hiponimia, antonimia, holonimia, etc.* (ver figura 4.8). Los nodos en la red semántica de WordNet se refieren a *synsets*. Cada *synset* representa un significado particular compartido por un grupo de palabras y frases. Cada palabra o frase en el *synset* es un sinónimo de otros con respecto a este particular sentido; y una palabra o frase con más de un sentido aparecerá en más de un *synset*. Los nodos *synsets*, después de todo, encarnan el concepto de sinonimia, mientras que otras relaciones se modelan como arcos entre los nodos (ver figura 4.9).

La información que aporta WordNet con sus relaciones léxicas y semánticas nos puede ayudar a predecir el significado de una palabra. Por ejemplo, la ocurrencia de la palabra "side" en el contexto de la palabra "bank" sugiere que bank podría clasificarse con el sentido "side of

[22] Relación de inclusión de unidades léxicas que va de lo más específico a lo más general.

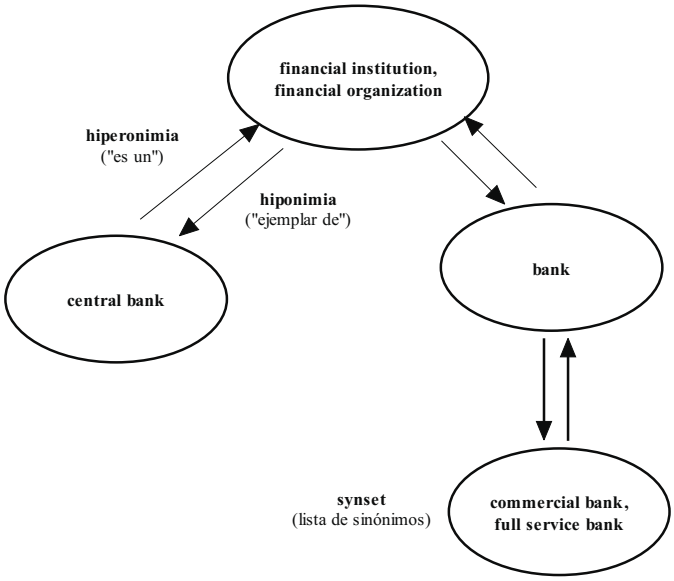

Figura 4.7: Fragmento de la relación jerárquica en WordNet

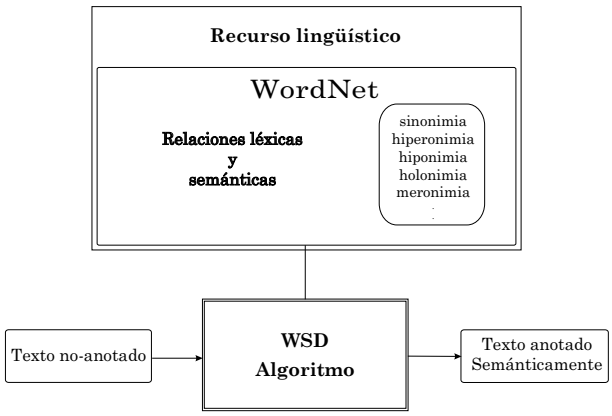

Figura 4.8: Arquitectura del enfoque WSD basado en base de datos léxica (WordNet)

a river" en vez de con el sentido "economical entity". El proceso de entrenamiento[23] se realiza como sigue. Se representa cada término de WordNet con un vector, conforme a lo relatado en el Modelo del Espacio Vectorial, cuyas componentes son: el peso del término y los pesos de los términos que constituyen la ventana contextual[24] para cada una de las relaciones mencionadas, si es que existen. Así, para cada uno de los términos del conjunto de entrenamiento, calcularemos su

[23]La totalidad de los nombres que se encuentran en WordNet ha constituido el conjunto de entrenamiento.

[24]Se construye una VC para cada significado del término, y para cada una de las relaciones que tenga el término dado. Así tendremos VC de sinónimos, hiperónimos, hipónimos, merónimos, etc. Lo que nos permitirá estudiar con qué tipo de relaciones se obtiene mayor *precision*.

ventana contextual, construyendo tantas ventanas como palabras con diferentes sentidos existan en la colección, así como relaciones existan para ese término. La fase de prueba se realiza confrontando la consulta, por medio del cálculo de la similitud con los vectores creados en el entrenamiento[25] con WordNet.

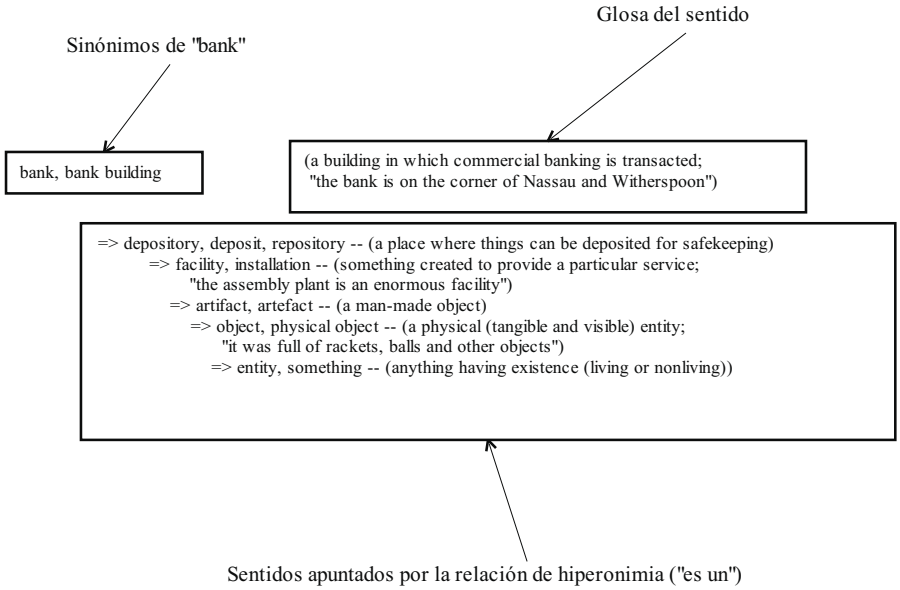

Figura 4.9: Ejemplo de información de sinonimia e hiperonimia extraída de WordNet para el término *bank*

Realizaremos una combinación de las distintas relaciones. Para llevar a cabo este proceso, proponemos un algoritmo desambiguador fundamentado en técnicas probabilísticas, concretamente un algoritmo *Monte Carlo p-correcto*, donde p es la precisión obtenida en los entrenamientos para cada una de las relaciones de WordNet propuestas. Al tomar la *precision* obtenida en el entrenamiento como probabilidad de acierto, la construcción del algoritmo es inmediata, ya que basta con hacer las desambiguaciones para todas las relaciones y ponderar los sentidos obtenidos con la *precision* correspondiente. De esta forma, tenemos en cuenta la aportación semántica de todas ellas, devolviendo el sentido más probable.

4.6.3 Desambiguador basado en la integración de recursos lingüísticos

Uno de los objetivos de este enfoque es la utilización de las relaciones léxicas y semánticas de WordNet para incrementar el tamaño del conjunto de entrenamiento de SemCor. Aunque la utilización de corpora de textos y bases de datos léxicas por separado ha producido avances interesantes en WSD, consideramos que los resultados pueden ser mejorados haciendo un uso

[25] Es decir, con las distintas ventanas contextuales.

integrado y combinado de ambos tipos de recursos. Partiendo de la hipótesis de que un sistema informático realiza su misión mejor si dispone de más información, nosotros planteamos un modelo integrador de WSD, orientado a emplear de manera combinada corpus de entrenamiento y bases de datos léxicas (ver figura 4.10). Así, al incorporar información proveniente de WordNet a la colección de entrenamiento, la efectividad de la desambiguación debe mejorar.

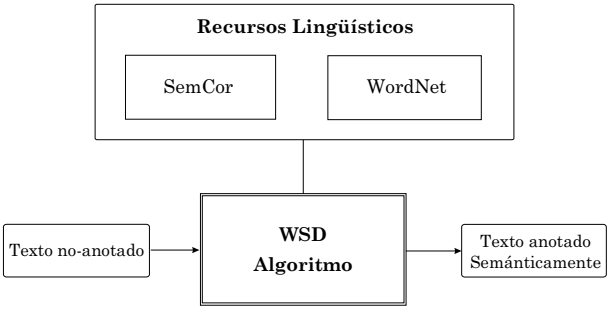

Figura 4.10: Arquitectura basada en la integración de recursos lingüísticos, corpus de entrenamiento y base de datos léxica

Realizamos una integración de recursos, utilizando información de sinonimia [Ureña et al., 1998c], hiponimia e hiperonimia, como se describe a continuación. En la fase de entrenamiento se construyen primeramente los vectores conforme a lo relatado en el enfoque basado en el entrenamiento, obteniendo un conjunto de vectores, uno por cada uno de los términos que constituyen la colección de entrenamiento. A continuación, cada uno de los términos que representan a los vectores, se consulta en WordNet, si dicho término tiene un *synset* asociado para el sentido consultado se "une" dicho *synset* al vector, recalculándolo con mayor peso, en caso contrario, se elimina dicho *synset*. De igual modo se hace con las relaciones de hiponimia e hiperonimia. La fase de prueba se realiza confrontando la consulta, por medio del cálculo de la similitud, con los vectores creados en el entrenamiento. Se obtiene una mejora en el proceso de desambiguación cuando se complementa el entrenamiento basado en corpus con información de sinonimia, hiponimia e hiperonimia.

Claramente identificamos el papel de cada recurso en este enfoque de desambiguación. Por un lado, WordNet amplia el número de términos en relación con un determinado sentido, cuando los datos de entrenamiento no son grandes o no son seguros. Esto directamente contribuye con los términos usados en la representación del vector. Por otro lado, la colección de entrenamiento proporciona mayor información contextual para aquellos términos mejor entrenados.

Para el algoritmo de Rocchio, nosotros hemos considerado el valor de proximidad semántica previamente producido como un número de ocurrencias de un término con un significado, así este valor es multiplicado por el peso del término en la colección. Por otro lado, la inserción del peso de un término para un significado en el algoritmo de Widrow-Hoff es normalizado por la constante η.

4.6.4 Estructura del desambiguador

En las figuras 4.11, 4.12 y 4.13 se incluyen los Diagramas de Flujo de Datos (DFDs) [Yourdon y Constantine, 1979; Rumbaugh, 1991; Pressman, 1997] correspondientes a los tres enfoques presentados del sistema WSD.

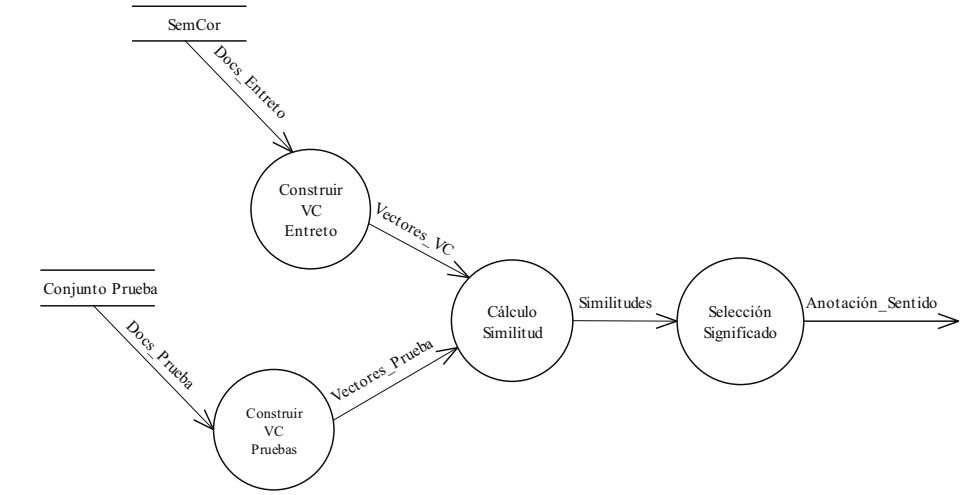

Figura 4.11: DFD que describe el enfoque de desambiguación basado en SemCor

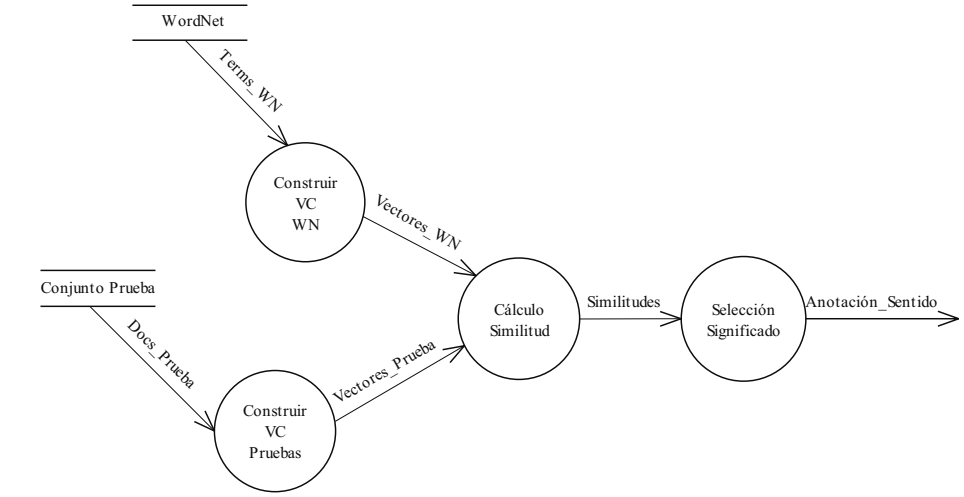

Figura 4.12: DFD que describen el enfoque de desambiguación basado en WordNet

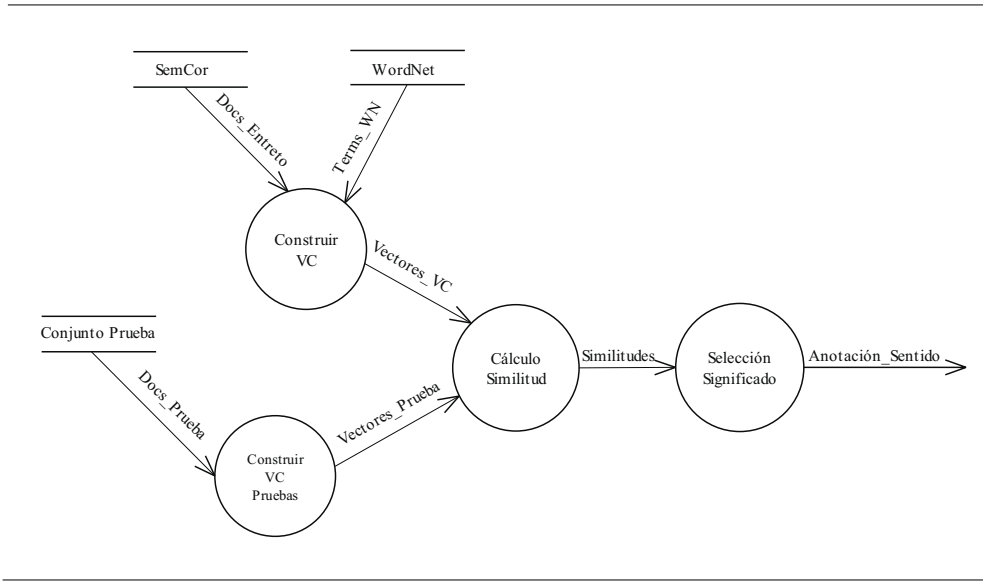

Figura 4.13: DFD que describe el enfoque de desambiguación basado en la integración de recursos lingüísticos

En los DFDs los elementos de datos del proceso de entrenamiento son:

- Para el desambiguador basado en SemCor (figura 4.11),

  - *Docs Entreto*, representa el conjunto de documentos de texto del corpus de entrenamiento SemCor.

- Para el desambiguador basado en WordNet (figura 4.12),

  - *Terms WN*, representa el conjunto de términos con sus significados y sus relaciones de la base de datos léxica.

- Para el desambiguador basado en la integración de recursos lingüísticos (figura 4.13),

  - *Terms WN*, que representa el conjunto de términos con sus significados y sus relaciones de WordNet, y
  - el elemento de datos *Docs Entreto*, que representa el conjunto de documentos de entrenamiento procedente del corpus de textos SemCor.

A continuación se describen los distintos elementos comunes a los tres enfoques.

El elemento de datos *Vectores VC* representa el fichero que se genera como consecuencia del proceso *Construir VC*, el fichero contiene las ventanas contextuales de entrenamiento de acuerdo con las técnicas antes detalladas, y que está formado por el conjunto de términos de indexación

s_{ji}, con sus pesos w_i, y la representación de los términos y sus significados mediante sus vectores $s_{ji} = \langle ws_{j1}, ws_{k1}, ..., ws_{kn} \rangle$.

Después, se describen los elementos de datos del proceso de prueba. El elemento de datos *Docs Prueba* proporciona los ficheros o documentos de prueba, estos documentos no están anotados semánticamente, y constituyen la entrada al sistema WSD. De manera análoga al proceso de entrenamiento, el elemento de datos *Vectores Prueba* representa el fichero que se genera como consecuencia del proceso *Construir VC Prueba*, el fichero contiene las ventanas contextuales de los documentos de prueba obteniéndose los términos de indexación c_k aparecidos en la consulta y los pesos asignados w_i, generándose el vector $c_k = \langle ws_{c1}, ws_{ck1}, ..., ws_{ckn} \rangle$.

Finalmente, se generan los elementos de datos *Similitudes*, que es un fichero que contiene los valores de similitud, después de confrontar los vectores de entrenamiento con la consulta. Y *Anotación Sentidos* que constituye la salida del sistema WSD, y que contiene el sentido seleccionado para un determinado término.

El proceso de entrenamiento (*Construir VC Entreto*) se ejecuta una vez, mientras que su ejecución tiene un mayor coste computacional que el resto de los procesos.

Los procesos de prueba (*Construir VC Prueba*) y de cálculo de similitud (*Calculo Similitud*) se ejecutan una vez por cada procesamiento de una consulta. Su ejecución es menos costosa computacionalmente. La salida del sistema se proporciona en un fichero. De esta forma puede ser utilizado por otras aplicaciones (IR, TC, etc.).

El sistema ha sido implementado íntegramente en *C* en el entorno Solaris 5.0 sobre una estación de trabajo Sun Enterprise Ultra 2.

4.7 Descripción del entorno experimental

En este punto presentamos el diseño de una serie de experimentos de WSD orientados al estudio de la efectividad de nuestros enfoques WSD. Para ello, hemos considerado dos cuestiones principalmente. En primer lugar, nos interesa comparar la efectividad de los enfoques con respecto a una aproximación línea base. En segundo lugar, estos experimentos también pretenden evaluar las mejoras producidas por la integración de recursos lingüísticos frente a la utilización de éstos por separado.

Nuestros experimentos centrados en la efectividad, se basan en medidas de los índices *recall* y *precision*.

4.7.1 Recursos empleados

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

Línea Base. Sólo se emplea para desambiguar la frecuencia de aparición de los sentidos en el corpus, asignando el sentido más frecuente. Este tipo de WSD constituye una línea comparativa básica para el resto de experimentos. A esta clase de experimentos la denominamos en este trabajo "WSD línea base".

Corpus de entrenamiento. Hemos seleccionado para esta experiencia la colección de documentos SemCor, y hemos denominado a estos experimentos "WSD basada en colección de entrenamiento."

Base de datos léxica. En este trabajo empleamos WordNet, y al experimento que hace uso de ella lo llamamos "WSD basada en WordNet".

Corpus de entrenamiento y base de datos léxica. Integrando y combinando la utilización de SemCor y WordNet pretendemos mejorar los resultados obtenidos que con el uso de ambos recursos por separado. Este experimento se denomina "WSD basada en WordNet y una colección de entrenamiento".

4.7.2 Tamaño de la ventana contextual

Uno de los objetivos de nuestros experimentos fue decidir entre diferentes unidades contextuales: frase, párrafo o bien otros tamaños de ventanas contextuales. Para ello, hemos seleccionado aleatoriamente cuatro documentos o textos de SemCor considerados individualmente: *br-a14, br-j09, br-k11* y *br-k14*. Estos textos han representado el papel de ficheros de entrada (sin etiquetas). Para comparar nuestro algoritmo basado en el corpus de entrenamiento SemCor, como hemos comentado anteriormente, hemos decidido implementar un algoritmo línea base [Boguraev y Pustejovsky, 1996], y confrontarlo con nuestros experimentos.

Hemos tomado diferentes tamaños de *ventanas contextuales* como adquisición de conocimiento (de 10 a 60 términos), y hemos realizado los experimentos para los mismos ficheros seleccionados anteriormente. Se muestra en la figura 4.14 la *precision microaveraging* y *macroaveraging*.

Podemos observar que, conforme aumenta el tamaño de la ventana contextual, mayor es la *precision* de nuestro algoritmo. Hemos considerado como tamaño máximo 60 términos circundantes, debido a que éste es el número más significativo computado en SemCor por nuestro algoritmo y, a partir de este tamaño se obtienen resultados similares. Se obtiene empíricamente que cuanto mayor es la ventana contextual, más *precision* alcanza nuestro algoritmo, puesto que se adquiere mejor contexto, así cuando el tamaño de la ventana contextual se aproxima al párrafo, tal y como lo define SemCor se consigue mayor *precision*.

Como se puede obsevar en la figura 4.15, la *precision microaveraging* y *macroaveraging* obtenida por nuestro algoritmo basado en SemCor, utilizando como tamaño de ventana contextual el párrafo, es mayor que la obtenida por el algoritmo línea base. La unidad contextual "párrafo" no tiene un tamaño fijo, sino variable en SemCor, puesto que cada párrafo, puede tener una o varias frases, con un número distinto de términos.

Concluyendo, utilizamos la ventana contextual de tamaño variable. En las figuras 4.16 y 4.17 se muestran ejemplos respectivos de ventanas contextuales construidas para el enfoque basado en SemCor y para el basado en WordNet. Como se puede observar, las ventanas contextuales constituyen el conjunto de entrenamiento, éste se encuentra ordenado alfabéticamente por el término representado. Los términos que integran la ventana contextual de un término dado,

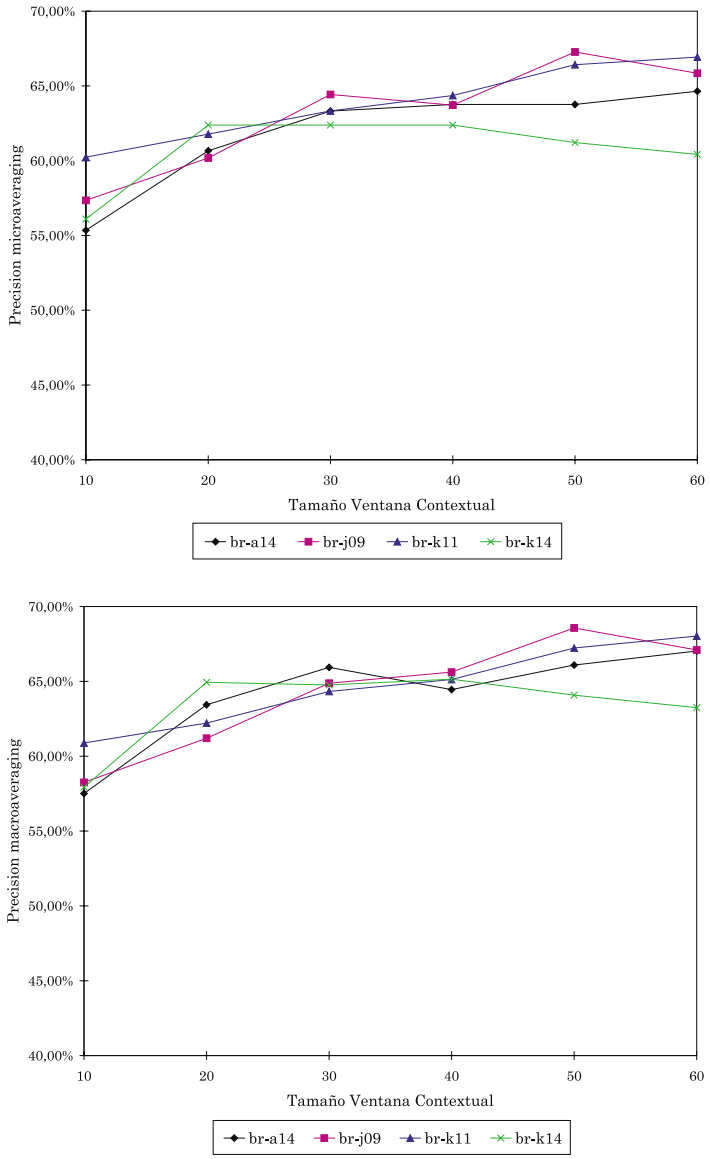

Figura 4.14: *Microaveraging* y *macroaveraging* para una ventana contextual con un tamaño en el intervalo [10,60]

también se encuentran ordenados alfabéticamente, incluyendo, para cada uno de ellos, el sentido y la frecuencia en dicha ventana. Así, por ejemplo en la figura 4.17, el término *bank* con sentido *7*, se representa como **bank#7**.

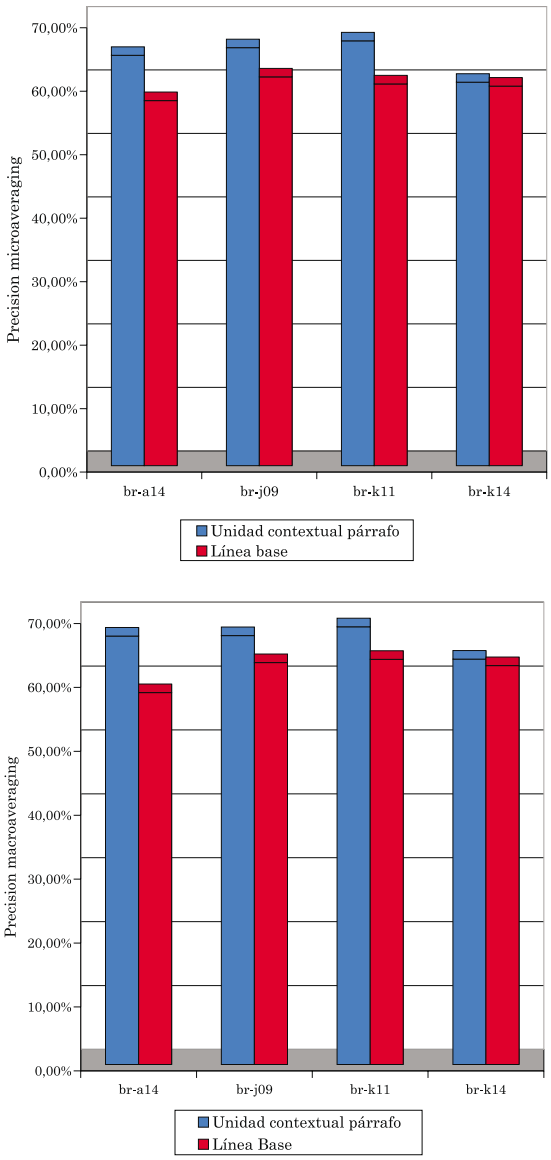

Figura 4.15: *microaveraging* y *macroaveraging* para una ventana contextual de tamaño "párrafo"

4.7.3 Resultados experimentales de la evaluación directa e interpretación

Describimos los experimentos realizados para probar el rendimiento y la efectividad de nuestros algoritmos de desambiguación. Utilizando una colección de prueba, los tres enfoques proporcionan los valores de similitud correspondientes. En los tres sistemas, los valores de similitud se utilizan como base para el ordenamiento de los sentidos por relevancia y selección del sentido con mayor valor de similitud.

No. 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 4.1: Porcentaje de ocurrencias de SemCor 1.6

bank#1 305 15\5#1 1 500\5#1 1 activity\1#1 1 advice\1#1 1 ago\4#1 1 aid\2#1 1 all\3#1 1 almost\4#1 1 alone\4#2 1 also\4#1 1 always\4#1 1 amount\2#1 1 annually\4#1 1 applicant\1#1 1 approval\1#1 1 approved\5#1 1 association\1#1 1 at_least\4#2 1 available\3#1 1 be\2#2 3 be\2#3 1 be\2#4 1 be\2#5 2 become\2#2 1 bedroom_set\1#1 1 bennington\1#1 1 better\4#1 1 bill\1#1 2 board_of_directors\1#1 1 bonnet\1#1 1 branch\1#1 1 build\2#1 1 business\1#1 3 businessman\1#1 1 candy_store\1#1 1 capital_stock\1#1 ...

....

bank#5 33 ballroom\1#1 1 below\4#4 1 best\3#1 1 black\1#1 1 black\3#1 1 buckle\1#1 1 careful\3#1 1 clear\2#3 1 come_down\2#1 1 come_in\2#1 1 dark\1#1 1 delayed\3#1 1 dressed\5#1 1 foot\1#7 1 forward\4#1 1 french_window\1#1 1 give_way\2#1 1 great\5#1 1 hold\2#12 1 jeweled\5#1 1 moon\1#1 1 moonlight\1#1 1 now\4#3 1 old\3#1 1 scud\2#1 1 shoe\1#1 1 silk\1#1 1 silver_gray\1#1 1 sky\1#1 2 slow\3#1 1 staircase\1#1 1 ...

...

bank#7 111 able\3#1 1 arc\1#2 1 arroyo\1#1 3 assembly\1#2 1 assume\2#1 1 automobile\1#1 1 begin\2#1 1 boston\1#1 1 brick\1#1 1 building\1#1 2 bullet\1#1 2 bypass\1#1 1 colored\3#1 1 comrade\1#1 1 concealed\5#1 1 conventional\3#1 1 cook\1#1 1 difficult\3#1 1 direction\1#1 1 ...

Figura 4.16: Ejemplo de Ventanas Contextuales construidas a partir de SemCor, para el término "bank" con los significados #1, #5 y #7

Las palabras de un documento tienen muy distintas frecuencias de aparición dentro de una colección [Sanderson, 1996]. Esto se puede demostrar examinando distintas colecciones, en concreto, vamos a estudiar esta cualidad en la colección SemCor, en relación con las palabras ambiguas. En la tabla 4.1 se muestra el porcentaje de ocurrencias con que aparecen los distintos sentidos. Así, los valores entre paréntesis indican el sentido más frecuente con el que aparecen en la colección, mientras que en la última columna, entre llaves, se muestra el porcentaje que resultaría si los sentidos apareciesen en igual cantidad, es decir, con una distribución proporcional para cada sentido.

De esta manera, estudiamos si son igualmente frecuentes todos los sentidos de una palabra ambigua. Por ejemplo, se puede ver cómo los términos que poseen 10 sentidos muestran en el 63% de los casos el sentido 1, por tanto, el resto de los sentidos no son igualmente probables ya que si existiese una distribución uniforme, cada sentido tendría una proporción del 10% en la colección. Por otra parte, también se puede apreciar en la tabla 4.1, que los términos polisémicos, independientemente del número de sentidos que puedan poseer, muestran con mayor probabilidad el sentido 1.

Aunque aproximadamente sólo el 18%[26] de las palabras en WordNet son polisémicas, en

[26]Sin embargo, este porcentaje representa aproximadamente a 15.400 formas de palabras polisémicas [Fellbaum, 1998].

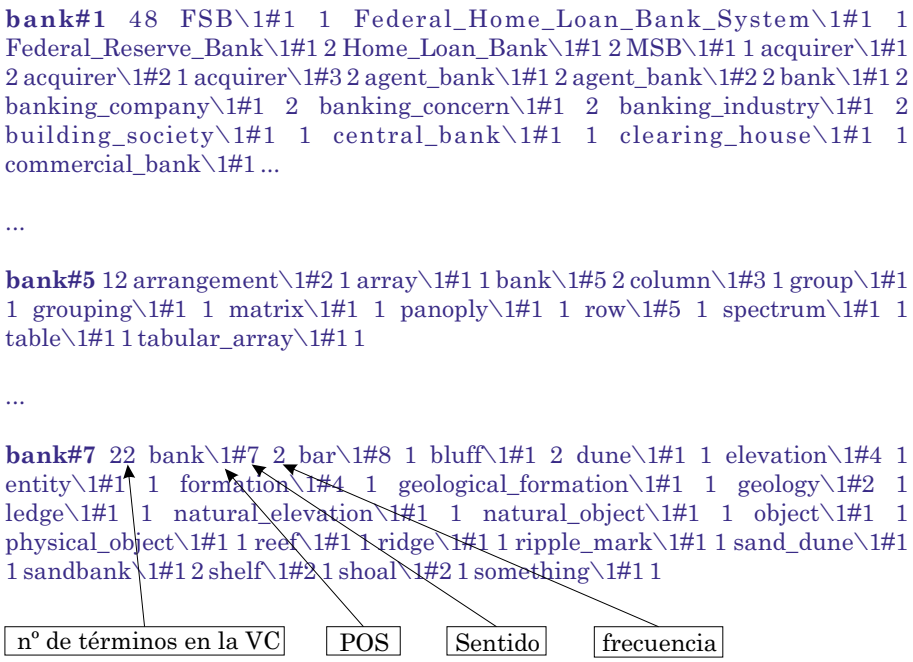

Figura 4.17: Ejemplo de Ventanas Contextuales construidas a partir de todas las relaciones de WordNet, para el término "bank" con los significados #1, #5 y #7

SemCor aproximadamente el 80% de las palabras son polisémicas. Esto se debe al simple hecho de que las palabras utilizadas frecuentemente tienden a tener diferentes sentidos.

Asimismo, se puede extraer de SemCor una gráfica que representa la distribución de sentidos en relación con la frecuencia de aparición de éstos (figura 4.18).

Algoritmo línea base

Consideraremos como algoritmo de comparación, un algoritmo base fundamentado también en el modelo del espacio vectorial de sistemas de recuperación de información [Salton et al., 1975]. Se construyen vectores para cada uno de los términos del corpus etiquetado, con tantos componentes como sentidos tenga. Los componentes de los vectores son pesos que reflejan la relativa importancia de los términos en el texto. Este enfoque con pesos fue diseñado para favorecer a aquellos términos que aparecen más frecuentemente con un determinado sentido [Leacock et al., 1996]. Los pesos se definen como:

$$w_s = p \cdot min(n_s, d) \tag{4.15}$$

Siendo:

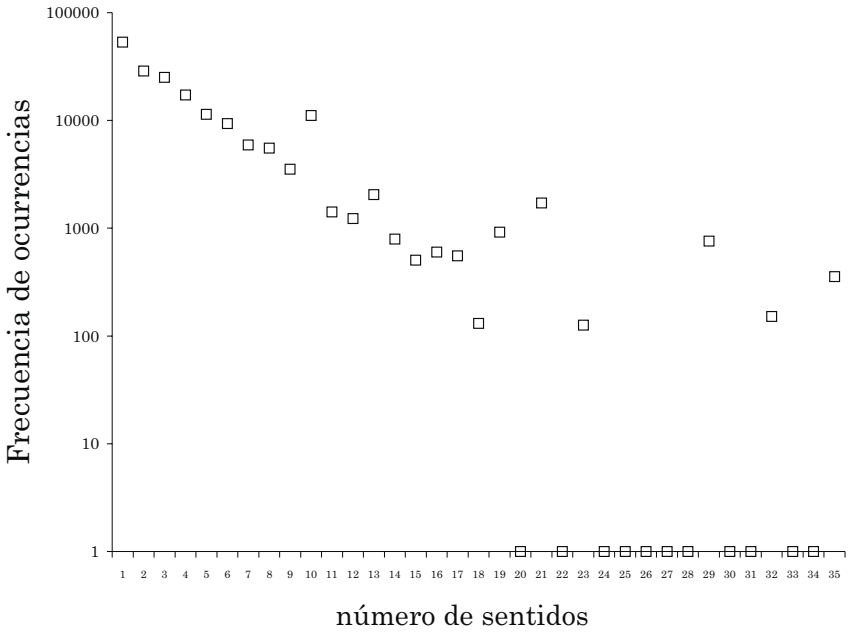

Figura 4.18: Distribución de sentidos en la colección SemCor 1.6 (escala logarítmica)

n_s	el número de ocurrencias de c con el sentido s
p	$= \frac{n_s}{\sum n_s}$
d	diferencia entre los dos valores mayores de n_s (si la diferencia es 0, d se establece a 1)

Así por ejemplo, si un término aparece seis veces con el sentido uno y ninguna vez con los cinco sentidos restantes, entonces el vector de entrenamiento contendrá seis componentes, cada uno con un peso, representado cada uno de los sentidos de la forma: $(6, 0, 0, 0, 0, 0)$. Sin embargo, un término que aparece 10, 4, 7, 0, 1 y 2 veces con sus respectivos sentidos, queda representado con sus pesos por medio del vector $(1.25, 0.5, 0.88, 0, 0.04, 0.17)$. En resumen, como se ha visto, este algoritmo es simple ya que asigna a todos los contextos el sentido más frecuente.

Enfoque basado en corpus

En la tabla 4.2 se muestran los resultados obtenidos en los experimentos realizados para el desambiguador basado en SemCor, para cuatro ficheros pertenecientes al *Brown1* elegidos aleatoriamente.

A diferencia de otros trabajos relevantes que para la evaluación del desambiguador se centran en seleccionar un conjunto reducido de palabras, generalmente con un par de sentidos de muy diferentes significados, nosotros probamos nuestro método con todos los nombres y verbos en

un subconjunto del corpus no restringido de dominio público, haciendo distinción entre el gran número de sentidos de WordNet, pudiendo observar los resultados tan próximos que obtienen ambos algoritmos.

	Rocchio		Widrow-Hoff		Línea base
docs	*micro*	*macro*	*micro*	*macro*	*precision*
a11, k10, k28, f10	60,56	62,26	67,50	62,44	56,38
Media	57,45	59,38	60,67	59,82	59,71

Tabla 4.2: Resultados de los experimentos para el enfoque basado en corpus de entrenamiento

La comparación de nuestros métodos, como hemos comentado anteriormente, se ha realizado con un algoritmo *línea base*. La *precision* obtenida por nuestro algoritmo es alta, como se puede observar en la tabla 4.2, donde se encuentran resumidos los resultados para nuestra primera serie de experimentos. Esta tabla muestra las medias *macroaveraging* y *microaveraging* de *precision*. No se presenta el *recall* por las razones aludidas en la sección 4.4, ya que este enfoque de desambiguación siempre toma una decisión por algún significado, con lo cual el *recall* es igual a 1 (cobertura del 100%). La desambiguación de todos los términos, cualquiera que sea la partición de la colección de entrenamiento, es un aspecto muy destacable de este enfoque.

Los resultados mostrados en la tabla, se han obtenido utilizando como colección de entrenamiento todos los documentos que integran el *Brown1*, exceptuando los documentos que serán evaluados posteriormente. La elección de los documentos de evaluación se ha realizado aleatoriamente sobre el *Brown1*, así en la primera fila de la tabla 4.2 se muestran los resultados de la evaluación realizada para los documentos *br-a11, br-k10, br-k28 y br-f10*. En la segunda, se muestra la *precision media* obtenida para la evaluación cruzada [Manning y Schütze, 1999] de todos los documentos del *Brown1*. Esta evaluación se ha realizado mediante el proceso de escoger al azar cuatro documentos, entrenando con todos los restantes y evaluando con los seleccionados. Este proceso se ha repetido hasta completar la evaluación con todos los documentos del *Brown1*.

	Rocchio		Widrow-Hoff	
precision	*micro*	*macro*	*micro*	*macro*
Media	56,12	58,46	58,72	58,47

Tabla 4.3: Evaluación realizada entrenando con el Brown2 y evaluando con el Brown1

En la tabla 4.3 se muestran los resultados de la evaluación del conjunto Brown1 para la totalidad de sus ficheros (103), bajo el entrenamiento del conjunto Brown2 (83 ficheros). Si comparamos los resultados obtenidos en las tablas 4.2 y 4.3, para los diferentes conjuntos de entrenamiento, observamos que los resultados medios obtenidos bajo el entrenamiento del Brown2 son levemente menores. Estas variaciones dependen del conjunto (partición) de entrenamiento y evaluación empleados.

En el Apéndice C sección C.1 se muestran detalles adicionales de los experimentos mediante una evaluación cruzada.

Enfoque basado en base de datos léxica

En la tabla 4.4 se muestran los resultados para cada una de las relaciones proporcionadas por WordNet. Podemos considerar como relaciones para nuestro enfoque las siguientes: sinonimia[27], hiponimia, hiperonimia, meronimia, holonimia, antonimia y coordinados, destacando la gran *precision* (*micro* y *macro*) para todas ellas.

Relaciones WordNet	Rocchio			Widrow-Hoff		
	micro	*macro*	*recall*	*micro*	*macro*	*recall*
ants	88,46%	91,43%	0,10%	93,33%	89,14%	0,12%
holon	79,38%	80,10%	0,38%	80,91%	80,29%	0,30%
hholn	78,26%	78,57%	0,45%	79,70%	77,83%	0,36%
hypo	65,23%	66,50%	3,01%	65,05%	66,04%	2,64%
tree	61,02%	62,16%	4,49%	61,31%	62,01%	4,30%
meron	60,14%	60,81%	0,56%	64,61%	60,82%	0,48%
hmern	59,53%	60,70%	0,84%	60,20%	60,65%	0,82%
hype	42,38%	43,21%	9,02%	43,06%	43,17%	7,52%
coorn	42,17%	46,15%	11,90%	43,41%	46,08%	10,22%
syns	25,44%	35,56%	3,57%	25,29%	35,67%	2,82%
Todas	50,37%	53,39%	21,46%	51,43%	52,93%	19,85%
Combinación relaciones	72,34%	73,95%	30,15%	76,11%	75,64%	30,61%

Tabla 4.4: *precision* y *recall* obtenidos por las distintas relaciones léxicas y semánticas del enfoque basado en WordNet en la evaluación del Brown1

Se ha empleado como colección de prueba para realizar la evaluación de este enfoque, el *Brown1* en su totalidad. Y como colección de entrenamiento se han utilizado todos los nombres y verbos de WordNet 1.6., construyendo tablas de entrenamiento independientes relativas a cada relación.

En los resultados de este método sí aparece un valor de *recall* para cada relación empleada, esto es como consecuencia de que este enfoque no decide en la totalidad de los casos. Por ejemplo, podemos observar la gran *precision* obtenida por este enfoque con la relación de *holonimia*.

Resaltar como este enfoque para cada una de las seis primeras relaciones de la tabla obtiene una *precision* superior al 60%, particularmente se puede resaltar la *precision* tan alta, del orden del 90%, que obtienen ambos algoritmos empleando la relación *antonimia*. Este valor tan alto sorprende y parece paradójico, pues cabría esperar mayor *precision* con otras relaciones más que

[27] La similitud de significados es la relación más importante de WordNet.

con la propia *antonimia*, sin embargo el *recall* es extremadamente bajo, sólo en el 0,1% de los casos aproximadamente es capaz de decidir.

En la tabla aparecen valores correspondientes a *"Todas"*, ésta no es una relación, sino la unión de todas ellas. Como podemos apreciar en esta unión la *precision* supera el 50%, y el *recall* aunque es superior comparado con cada una de las relaciones, es relativamente bajo (21%).

La figura 4.19 nos da una visión general muy clara del comportamiento para cada una de las relaciones de WordNet. Se puede observar que los mejores valores de *recall* son los obtenidos por las relaciones de hiperonimia (hype) y coordinados (coorn), lo que provoca una pequeña disminución de *precision*.

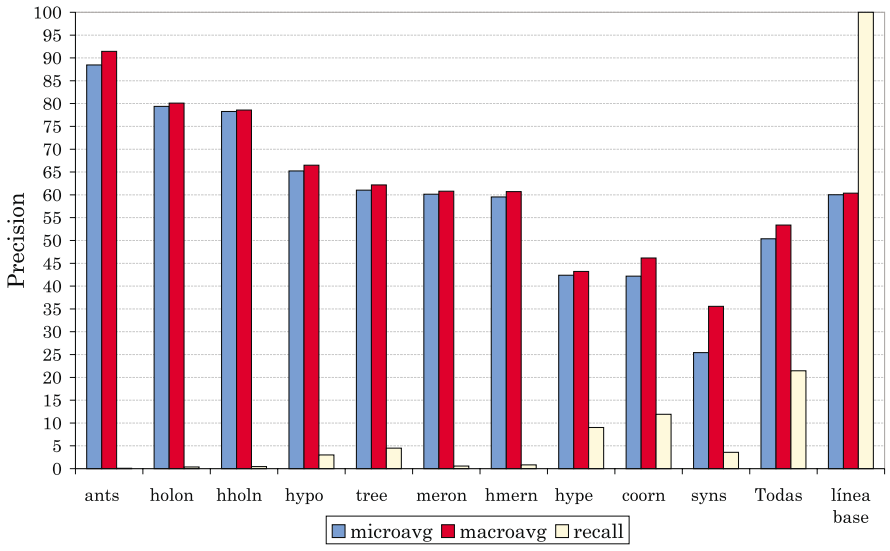

Figura 4.19: Relación entre *precision* (*microaveraging* y *macroaveraging*) y *recall* basada en las relaciones de WordNet

Combinar los resultados mostrados en la tabla 4.4 puede llevarnos a mejorar el desambiguador. Podemos ver la *precision microaveraging* como una probabilidad de acierto en la desambiguación con las relaciones léxicas y semánticas. Si hacemos un mismo experimento para todas las relaciones, y ponderamos los resultados según dichas relaciones, obtendremos la probabilidad de cada uno de los sentidos posibles. Supongamos que un término t_i se intenta desambiguar mediante varias relaciones, asignando cada una de ellas los sentidos siguientes: t_{i1} *(ants)*, t_{i2} *(holon)*, t_{i1} *(hypo)*, t_{i4} *(meron)*, t_{i1} *(hype)*, t_{i-} *(coorn)* y t_{i2} *(syns)*.

A priori se ve que el sentido 1 es el más probable, aunque éste depende de las probabilidades de las relaciones. Para el cálculo usaremos un algoritmo probabilístico, en particular *Monte Carlo*, ya que podemos integrarlo directamente en nuestro desambiguador. La respuesta del algoritmo será un vector con las probabilidades de cada uno de los sentidos posibles. El cálculo

se hará en sentido decreciente de la *precision microaveraging* de cada una de las relaciones, y la respuesta será acumulada en la posición n-ésima del vector correspondiente a la relación en curso, atendiendo a la fórmula:

$$a_n^t = a_n^{t-1} + (1 - a_n^{t-1}) \cdot microavg_n/100 \tag{4.16}$$

Donde t es el instante de cálculo y n la relación de WordNet empleada.

De acuerdo con lo anterior, se obtienen los resultados de la tabla 4.5.

				vector de resultados			
paso t	relación	microavg	sentido	n=1	n=2	n=3	n=4
1	ants	88,46	1	0,8846	0,0000	0,0000	0,0000
2	holon	79,38	2	0,8846	0,7938	0,0000	0,0000
3	hypo	65,23	1	0,9599	0,7938	0,0000	0,0000
4	meron	60,14	4	0,9599	0,7938	0,0000	0,6014
6	hype	42,38	1	0,9769	0,7938	0,0000	0,6014
5	coorn	42,17	-	0,9769	0,7938	0,0000	0,6014
7	syns	25,44	2	0,9769	0,8463	0,0000	0,6014

Tabla 4.5: Ejemplo de combinación de relaciones

Las relaciones que se han combinado mediante la técnica *Monte Carlo* son: antonimia, holonimia, hiponimia, meronimia, hiperonimia, coordinados y sinonimia, mostradas en la tabla 4.4. Los resultados finales de la combinación de las citadas relaciones mediante el enfoque *Monte Carlo* muestran un buen valor de *precision* (mayor del 75% para el algoritmo de Widrow-Hoff), y un incremento del *recall* en relación con las relaciones.

Por último, a la vista de los resultados, podemos afirmar que, WordNet es una herramienta de muy buena calidad para la desambiguación y discriminación de los sentidos de las palabras, aportando información tremendamente relevante con el contexto que hemos definido. Sin embargo, no es perfecta, ya que no cubre la totalidad de los contextos, por lo que estamos limitados en *recall*, y por tanto, este enfoque de desambiguación combinando la relaciones no es capaz de decidir en el 70% de los casos aproximadamente.

En el Apéndice C sección C.2 se muestran detalles adicionales de los experimentos basados en WordNet. Se exponen los 50 primeros documentos del Brown1, evaluados para cada una de las relaciones léxicas y semánticas presentadas.

Enfoque basado en la integración de recursos lingüísticos

En la tabla 4.6 se muestra la integración de SemCor con información de WordNet. Como se puede observar al incorporar mayor información al entrenamiento aumenta la *precision*. El entrenamiento se ha realizado con la totalidad de los documentos pertenecientes al *Brown2*,

Recurso Lingüístico	Rocchio		Widrow-Hoff	
	microavg	*macroavg*	*microavg*	*macroavg*
SemCor Brown2	55,81%	58,02%	58,87%	57,93%
SemCor+WordNet Brown2	60,67%	63,06%	63,47%	64,55%

Tabla 4.6: Resultados obtenidos entrenando con el *Brown2* e integrando el entrenamiento con información de sinonimia de WordNet

donde cada término del entrenamiento ha sido expandido con información de sinonimia, hiperonimia e hiponimia de WordNet. Como colección de prueba para los experimentos expuestos en la tabla 4.6 se han empleado todos los documentos del *Brown1*, con lo que podemos afirmar que hemos realizado una evaluación completa y exhaustiva. En la tabla 4.6 se presenta en la primera fila (SemCor) los resultados para el enfoque basado en corpus, y en la última (SemCor+WordNet) los resultados después del proceso de integración de información de la base de datos léxica con el corpus.

A pesar de los significativos resultados obtenidos por el enfoque basado en WordNet, y en particular para determinadas relaciones, veíamos que el principal inconveniente era el valor tan pequeño de *recall* (30%) mediante la combinación. Sin embargo, con la integración de recursos lingüísticos se produce un incremento en *precision* de un 9% aproximadamente, con lo que la *precision microaveraging* asciende a más del 63% con el algoritmo de Widrow-Hoff, manteniéndose el *recall* total.

Por otra parte, podemos combinar mediante un algoritmo *Monte Carlo* el enfoque basado en la integración de recursos lingüísticos y en las relaciones de WordNet, al igual que se ha realizado para este último enfoque en la sección anterior. En la tabla 4.7 se presenta el entrenamiento basado en corpus con información de sinonimia, así como los resultados correspondientes a determinadas relaciones de WordNet para su combinación. Mediante la combinación del enfoque integrado y las relaciones, la *precision* alcanza el 70%, siendo el *recall* de un 55% aproximadamente.

Concluyendo, hemos presentado un enfoque robusto y eficiente para la resolución de la ambigüedad léxica integrando dos recursos lingüísticos con una *precision* próxima al 70%.

4.8 Resumen y conclusiones

En este capítulo, primeramente hemos descrito la tarea WSD y su terminología, y hemos hecho una revisión del estado del arte en desambiguación.

Relaciones WordNet	Rocchio			Widrow-Hoff		
	micro	*macro*	*recall*	*micro*	*macro*	*recall*
SemCor+WordNet Brown2	60,67%	63,06%	100%	63,47%	64,55%	100%
ants	88,46%	91,43%	0,10%	93,33%	89,14%	0,12%
holon	79,38%	80,10%	0,38%	80,91%	80,29%	0,30%
hypo	65,23%	66,50%	3,01%	65,05%	66,04%	2,64%
tree	61,02%	62,16%	4,49%	61,31%	62,01%	4,30%
meron	60,14%	60,81%	0,56%	64,61%	60,82%	0,48%
hype	42,38%	43,21%	9,02%	43,06%	43,17%	7,52%
coorn	42,17%	46,15%	11,90%	43,41%	46,08%	10,22%
syns	25,44%	35,56%	3,57%	25,29%	35,67%	2,82%
Combinación relaciones	65,07%	67,89%	55,25%	70,63%	70,70%	56,06%

Tabla 4.7: *precision* y *recall* obtenidos por la combinación de recursos lingüísticos en la evaluación de la colección de prueba *Brown1* de SemCor

Se han presentado los aspectos centrados en la efectividad como los más relevantes para nuestro trabajo. Hemos tratado la utilización de los índices *recall* y *precision*, así como los de *microaveraging* y *macroaveraging*, como los más adecuados para el estudio de la efectividad de los sistemas.

Mostramos especial interés por la integración de recursos lingüísticos, en particular por los corpora de textos y las bases de datos léxicas, al mejorar la efectividad. Damos un paso hacia métodos de análisis del contenido textual más avanzados e inteligentes. De acuerdo con lo anterior, se ha propuesto un modelo para la resolución de la ambigüedad léxica de las palabras basado en la integración de recursos lingüísticos, con el objetivo de poder a aplicarlo a tareas de clasificación automática de documentos. El modelo está uniformemente fundamentado en el modelo del espacio vectorial utilizado en recuperación de información, en el que existe una metodología de estudio experimental.

Se han desarrollado una serie de experimentos para el estudio de la efectividad de los modelos propuestos. En estos experimentos se ha utilizado una colección de prueba, el *Brown1* de SemCor en su totalidad para realizar la evaluación automática de manera minuciosa.

Asimismo, el enfoque basado en corpus de entrenamiento integrándolo con información lingüística se presenta como decisiva para esta mejora en la efectividad respecto a una aproximación basada simplemente en corpus de entrenamiento, o en base de datos léxica.

Capítulo 5

Aplicación de WSD a tareas de clasificación de documentos

5.1 Introducción

Durante los últimos años ha aumentado el interés por determinadas tareas propias de los sistemas de clasificación automática de documentos o más recientemente llamadas tareas de acceso a la información[1], al ayudar y permitir a los usuarios acceder a gran cantidad de información textual disponible en Internet y en sus organizaciones. Como se ha visto, estos sistemas realizan diferentes operaciones de clasificación, basándose en el análisis del contenido de los textos que procesan.

La clasificación de documentos puede mejorar con el empleo de recursos lingüísticos al permitir un análisis del contenido textual más rico. La tarea de resolución de la ambigüedad puede constituir un primer paso en este análisis, siendo posible la aplicación de esta tarea básica e intermedia, a otras más concretas de clasificación automática de textos, como son la recuperación de textos y categorización, donde la ambigüedad de significados constituye un problema considerable[2]. Estudiamos como con la desambiguación se puede mejorar el proceso de integración de recursos lingüísticos tanto en la recuperación de información, como en la categorización textos.

En este capítulo se trata la aplicación de la resolución de la ambigüedad léxica a dos tareas de clasificación. Primeramente, en la recuperación de textos se aplica la desambiguación (utilizando la técnica de realimentación) a la expansión de las consultas, basándose en la constatación de la dificultad de formular consultas que se muestren efectivas recuperando información relevante. En segundo lugar, para la categorización de textos se describe un enfoque basado en la integración de una colección de entrenamiento (Reuters-21578) y la base de datos léxica WordNet (versión

[1] El acceso a la información ha sido definido como un conjunto de tareas que facilitan dicho proceso, las cuales han llegado a ser el centro de muchos de los esfuerzos en investigación en este área. En concreto, la tarea de recuperación de texto es el objeto de la Conferencia TREC (Text REtrieval Conference) [Harman y Voorhees, 1997].

[2] Sanderson [1996] expone ejemplos de los efectos reales producidos por la ambigüedad en un sistema de recuperación de información on-line.

1.6) como fuentes de conocimiento. La información procedente de WordNet debe ser filtrada para hacer un uso efectivo de ella en el modelo de TC. Este proceso de filtrado es la aplicación de la desambiguación a la tarea de categorización.

La organización de este capítulo es como sigue. Comenzamos con la descripción de la aplicación de la desambiguación a la tarea de recuperación de información en el proceso de expansión de consultas. Después, se describe el entorno de evaluación en el que se exponen los resultados de los experimentos. A continuación, se estudia la categorización de textos haciendo uso de la integración de recursos lingüísticos, y se describe el proceso de aplicación de la desambiguación a esta tarea. Seguidamente se exponen los experimentos realizados. El capítulo termina con un resumen y conclusiones.

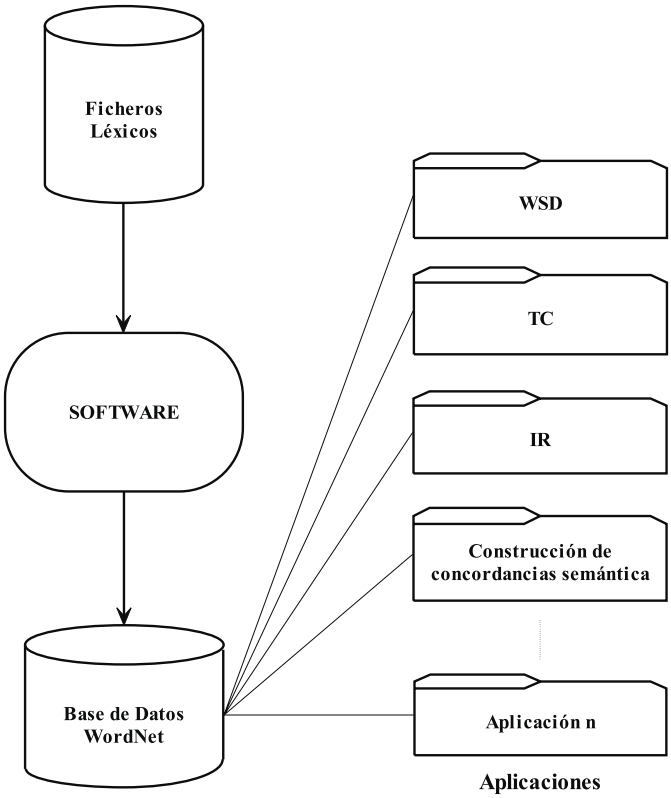

Figura 5.1: El sistema WordNet y aplicaciones

5.2 Aplicando la desambiguación a la Recuperación de Información

Los recursos lingüísticos en recuperación de información pueden representar un papel importante, tanto en un entorno monolingüe como multilingüe, con bases de datos léxicas en uno o varios

idiomas.

Un sistema de IR está condicionado por las características del texto, una de éstas es la ambigüedad de los términos. La ambigüedad de éstos puede reducir o limitar la efectividad del proceso de recuperación. Puesto que muchas palabras son ambiguas en algún grado, la aplicación de la desambiguación a la recuperación de información va a permitir mayor transparencia a muchos usuarios, en cuanto al proceso de formulación de consultas. Asimismo, podrá disminuir la dificultad de interacción con el sistema, ya que las consultas podrán ser formuladas de manera más flexibles y más cercanas al lenguaje natural.

Estudios previos de resolución de la ambigüedad léxica en recuperación de información afirman que, el problema al que se enfrentan, es la escasa información de la consulta (información contextual) para poder desambiguar los términos que aparecen en ella. Nuestro modelo de desambiguación necesita información contextual para poder resolver la ambigüedad de los términos, por lo que su aplicación a la recuperación requiere información adicional para poder desambiguar las palabras que el usuario introduce en la consulta. Para solventar este problema nos hemos centrado en el proceso de recuperación que incluye la técnica de realimentación. Utilizamos los documentos que el usuario identifica como relevantes como información contextual para así poder desambiguar los términos de la consulta.

5.2.1 Procesamiento de la consulta

En un sistema de recuperación de información típico la consulta del usuario es el único mecanismo para averiguar sus necesidades de información. Los sistemas de IR utilizan las palabras de la consulta del usuario para identificar los documentos más sensibles a la consulta dada, clasificados por orden decreciente de relevancia[3].

Incluso hasta los mejores sistemas de IR tienen limitado el *recall* [Harman, 1992b]; los usuarios recuperan pocos documentos relevantes en respuesta a sus consultas, pero casi nunca todos los documentos relevantes. En aquellos casos en que el *recall* es crítico los usuarios raras veces tienen otros caminos para recuperar los documentos relevantes. Frecuentemente las consultas contienen términos que no aparecen en los ficheros de indexación de la mayoría de los documentos relevantes. Y casi siempre los documentos relevantes no recuperados están indexados por un conjunto diferente de términos. Este problema plantea la principal dificultad de los sistemas de información [Lancaster, 1969]. Los sistemas de IR utilizando solamente la consulta inicial están limitados en cuanto al número de documentos relevantes que son capaces de recuperar, por lo que es necesario modificar la consulta inicial para obtener mejores resultados [Rijsbergen, 1986].

En su trabajo van Rijsbergen propone la modificación de la consulta como recurso para aumentar el rendimiento de la búsqueda. La solución a este problema pasaría, entonces, por "expandir" la consulta, es decir, añadir otras palabras que permitieran recuperar más documentos relevantes sin incluir demasiada información irrelevante. Para modificar la consulta se han venido utilizando con éxito dos técnicas, la realimentación por relevancia (*relevance feedback*) que se basa

[3]Utilidad de la información para el usuario de acuerdo con su consulta.

en el ajuste de los pesos de los términos de la consulta, y la utilización de recursos lingüísticos como diccionarios, thesaurus o bases de datos léxicas, mediante los cuales se incrementan los términos de la consulta.

5.2.2 La tarea de recuperación con realimentación por relevancia

El proceso de realimentación por relevancia fue introducido en la mitad de la década de los 60, como un proceso automático para la reformulación de la consulta y así incrementar el rendimiento. Convencionalmente la formulación de la consulta o el proceso de reformulación ha sido una tarea manual, o más bien intelectual. Como se ha comentado, la realimentación por relevancia es una forma de expansión de la consulta, donde la idea fundamental consiste en seleccionar los términos o expresiones importantes, unidos a algunos documentos recuperados que son identificados como relevantes por el usuario, y formular una nueva consulta subrayando la importancia de esos términos.

Por tanto, esta técnica consiste en utilizar el juicio del usuario sobre la relevancia o no de los documentos recuperados a fin de mejorar la efectividad de la consulta inicial. Esta mejora se consigue de dos maneras: ajustando los pesos de los términos de la consulta y añadiendo nuevos términos a la misma. El estudio de la distribución de cada término en los documentos relevantes e irrelevantes permite aumentar o disminuir su peso, o lo que es lo mismo, su importancia relativa. Además, si por ejemplo, todos los documentos que el usuario juzga como relevantes contienen un determinado término, entonces puede tratarse de un buen término para añadir a la consulta original [Salton y McGill, 1983].

Por un lado, la realimentación por relevancia aísla al usuario de los detalles del proceso de formulación de la consulta, permitiendo la construcción de declaraciones de búsqueda útiles sin necesidad de tener un gran conocimiento de la colección, ni del entorno de búsqueda[4]. Y por otro, proporciona un proceso de alteración de la consulta controlado, diseñado para subrayar la importancia de algunos términos y desestimar otros.

La realimentación por relevancia releva al usuario de búsquedas intermedias, generando automáticamente nuevas formulaciones de consultas basadas en la valoración de relevancia de los usuarios durante las operaciones de búsqueda iniciales. La consulta construida nuevamente tendrá una similitud alta con el conjunto de documentos previamente identificados como relevantes, y una similitud baja con el conjunto de documentos no relevantes. Asumiendo que el conjunto de documentos relevantes D_R con respecto a una consulta es conocido, y también el conjunto de documentos no relevantes D_{N-R}, la mejor consulta Q es aquella que maximiza la función F, definida como la diferencia entre la similitud media consulta-documento para todos los ítems relevantes y la similitud media consulta-documento para los no relevantes.

[4]Cuando la modificación de la consulta se lleva acabo manualmente, el proceso es difícil de controlar, primero porque las características de los ítems relevantes y no relevantes no son conocidas perfectamente, y también, porque las características de los documentos no son transformables fácilmente a formulaciones de consulta correctas.

$$F = \overline{SIM}(Q, D_i)_{D_i \in D_R} - \overline{SIM}(Q, D_i)_{D_i \in D_{N-R}} \tag{5.1}$$

$\overline{SIM}$ representa el coeficiente de similitud medio entre la consulta y el conjunto de todos los documentos incluidos en los subconjuntos de documentos relevantes y no relevantes respectivamente. Cuando la similitud entre vectores se mide por el coseno, la consulta óptima tiene el peso del término proporcional a la diferencia entre la media de los pesos de los términos en los relevantes y la media de los pesos en los ítems no relevantes.

Un fórmula típica usada para construir una consulta mejorada Q' a partir de una consulta original Q, viene dada por:

$$Q' = \alpha(Q) + \beta \left(\frac{1}{R'} \sum_{D_i \in D_{R'}} D_i \right) - \gamma \left(\frac{1}{N'} \sum_{D_i \in D_{N'}} D_i \right) \tag{5.2}$$

Donde R' es un conjunto de documentos identificados previamente como relevantes, y N' un conjunto de documentos identificados como no relevantes. α, β y γ son constantes, y los términos entre paréntesis representan la consulta inicial, y la media de los documentos identificados previamente como relevantes y no relevantes.

Teniendo en cuenta que la diferencia entre la media de documentos relevantes y no relevantes es equivalente a la distancia entre los vectores correspondientes en el espacio vectorial. La efectividad del proceso de recuperación muestra, tanto con pruebas experimentales como formales [Salton y McGill, 1983], que el proceso de realimentación por relevancia es:

- más efectivo, cuando tanto los documentos relevantes se encuentran muy agrupados entre ellos (por eso la similitud es mayor), como los no relevantes. Y además si la distancia entre ambos grupos (entre los ítems relevantes y no relevantes) es lo más larga posible.
- menos efectivo, en el caso más realista, cuando el conjunto de documentos no relevantes cubren un área amplia del espacio. La distancia correspondiente entre relevantes y no relevantes es más pequeña.
- muy desfavorable, cuando los documentos relevantes y no relevantes están entremezclados.

Como decimos la realimentación por relevancia es una técnica muy utilizada en IR[5]. Los experimentos de Salton y Buckley [1990] mostraron que los métodos de realimentación *Ide dec-hi* [Ide, 1971] obtenían los mejores resultados para el modelo del espacio vectorial. Este método deriva el vector de la nueva consulta a partir de la consulta inicial, todos los documentos considerados relevantes y sólo el mejor clasificado de los documentos no relevantes. Más formalmente, siendo Q_0 el vector de la consulta original, R_i el vector del documento relevante i y S el vector del documento no relevante mejor clasificado. Se muestra en 5.3 como calcular el vector Q_1 correspondiente a la consulta expandida según este método.

[5]También la realimentación se ha empleado en otras tareas de análisis del contenido con resultados significativos [Maña et al., 2000].

$$Q_1 = Q_0 + \sum_{relevantes} R_i - 1/2S \tag{5.3}$$

Los documentos relevantes e irrelevantes cuyos vectores se utilizan en 5.3 son aquellos documentos recuperados al utilizar la consulta inicial sobre los que el usuario emite una valoración. Sin embargo, cuando se evalúa la realimentación mediante colecciones de prueba (esto es, colecciones que incluyen un corpus de documentos, un conjunto de consultas y los juicios de relevancia correspondientes) no se necesita usuarios. Se supone que las valoraciones se realizan sobre los n primeros documentos recuperados. La relevancia de cada documento se obtiene de la propia lista de juicios de relevancia proporcionada por la colección de prueba. Para nuestros experimentos hemos fijado el valor de n a 15.

Para evaluar el incremento en efectividad producido por el método de realimentación, no se puede simplemente, comparar los resultados obtenidos por la consulta inicial y la consulta expandida. Parte de la mejora estaría provocada por el aumento de posiciones en la lista de resultados de los documentos utilizados en la realimentación y ya vistos por el usuario. Para evitar esta distorsión hemos utilizado el método de la colección residual [Chang et al., 1971] que supone que el usuario inspecciona los n primeros documentos de la lista de resultados para emitir su valoración. El resto de documentos, que recibe el nombre de colección residual, se utiliza para volver a medir la efectividad usando la consulta inicial y la expandida mediante realimentación.

5.2.3 Utilización de WordNet en la recuperación

Como se ha visto, el proceso de recuperación de texto trata de recuperar todos los documentos que son relevantes en una colección para una consulta dada. Al igual que hemos utilizado un recurso lingüístico a escala real como es WordNet para la resolución de la ambigüedad, podemos considerar que también tiene un gran potencial para la recuperación de información, ya que proporciona palabras semánticamente relacionadas. Por ejemplo, los términos *plant, flora, plant life* se pueden identificar como ocurrencias del mismo concepto *"a living organism lacking the power of locomotion"*, frente a *plant, works, industrial plant* que se identifican como ocurrencias de otro concepto *"building for carrying on industrial labor"*. Se puede utilizar la sinonimia en WordNet para medir la distancia semántica entre ocurrencias de términos y así obtener maneras más sofisticadas de comparar documentos y consultas.

En tareas de clasificación de texto como la categorización, veremos en la sección 5.3.1 la utilización de la relación de sinonimia de WordNet para incrementar la cantidad de información de la que hace uso el sistema con resultados significativos [Ureña et al., 2001]. Así pues, en recuperación, la expansión de consulta con WordNet ha mostrado ser potencialmente relevante para aumentar el *recall*, porque permite recuperar documentos relevantes que podrían no contener los términos de la consulta [Smeaton et al., 1995]. Sin embargo, Sanderson [1996] estudia el problema de la desambiguación haciendo uso de *pseudopalabras* en la recuperación de información, estimando que si la desambiguación no se lleva a cabo con al menos el 90% de *precision*, los resultados que se obtienen pueden ser hasta peores que no desambiguando.

	Documentos corpus WSJ	consultas TREC
Cantidad	5.000	50
N. términos		
máximo	403	9
mínimo	1	1
media	21,664	3

Tabla 5.1: Número de términos contenidos en el corpus WSJ y en las consultas TREC

Una combinación de técnicas basadas en WordNet se propuso en el trabajo de Richardson y Smeaton [1995], en las que se incluían entre otras la desambiguación automática, mostrando unos resultados poco satisfactorios. Sin embargo, en [Smeaton y Quigley, 1996] se muestra una razonable mejora en la recuperación de texto sobre una pequeña colección de pies de imágenes (documentos muy cortos) utilizando medidas de distancia conceptual entre palabras basadas en WordNet 1.4. Previamente se desambiguaban manualmente las consultas y los pies de la imágenes.

Por otra parte, en [Gonzalo et al., 1998a] se propone un enfoque de recuperación de texto fundamentado en el modelo del espacio vectorial y basado en la indexación en términos de los *synsets* de WordNet en lugar de sus palabras. Mostrando que dicha indexación puede ayudar a mejorar la recuperación de texto.

La expansión de la consulta con WordNet ha sido utilizada en otros trabajos de recuperación de información [Voorhees, 1994; Harman, 1993]. Por ejemplo, Voorhees expandió manualmente 50 consultas TREC utilizando información de WordNet (1.3), encontrando que dicha expansión era útil en consultas cortas e incompletas. En general, los resultados de expansión de consultas haciendo uso de la desambiguación han sido poco satisfactorios, debido fundamentalmente al tamaño de la consulta, que suele ser pequeño (3 términos de media en la colección TREC-1, como se puede ver en la tabla 5.1), con lo cual hay una imposibilidad de resolver adecuadamente la ambigüedad al carecer de suficiente información contextual.

Proponemos el uso de la realimentación en el proceso de desambiguación de términos en recuperación de información mediante el empleo de recursos lingüísticos.

Incorporamos la técnica descrita de realimentación a nuestro enfoque de desambiguación basado en la integración de recursos lingüísticos, tratado en la sección 4.6.3, para solventar la escasa información contextual de las consultas, debido al tamaño de las mismas (ver figura 5.1). Realizamos una expansión de la consulta basada en la relación de sinonimia de WordNet, haciendo uso de la desambiguación [Ureña et al., 2000a,b].

El proceso es como sigue, se realiza la consulta y se obtiene una relación de documentos ordenados por orden de relevancia. A continuación, se utiliza el juicio del usuario, es decir aquellos documentos relevantes, para realimentar la consulta original y así poder desambiguar y expandir los términos de la consulta original con los *synsets* adecuados de WordNet (ver figura

5.2).

A continuación, se muestra detalladamente el proceso de recuperación basado en la realimentación y expansión de la consulta con información de sinonimia de WordNet utilizando WSD:

1. Formular la consulta
2. Obtener los documentos más próximos
3. Someterlos a juicio del usuario
   - Relevantes
   - Irrelevantes
4. Expandir la consulta
   - Aplicar WSD a la consulta original para integrar información de sinonimia de WordNet
   - Añadir los términos más relevantes de los documentos clasificados como tales
5. volver a la consulta

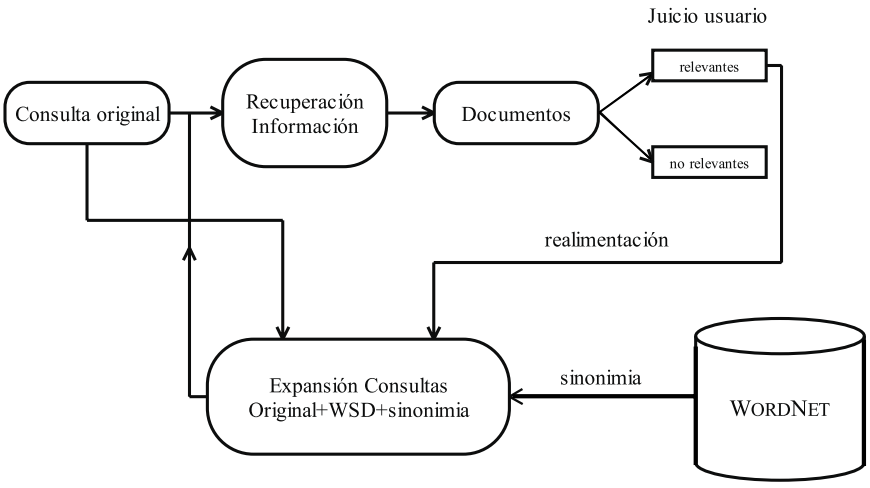

Figura 5.2: Proceso de recuperación de información con expansión de la consulta

En un último conjunto de experimentos hemos expandido las consultas TREC empleando la relación de sinonimia proveniente de WordNet. Hemos realizado el proceso de expansión de dos formas distintas. A continuación indicamos las características de ambas:

1. Para cada palabra de la consulta se toman todos los términos correspondientes a todas las categorías gramaticales con las que aparece y todos los significados posibles.

2. Partiendo de la colección de sinónimos anterior se realiza una desambiguación automática. Para ello, se utiliza el desambiguador basado en la realimentación.

En la figura 5.3 se muestran los distintos sinónimos de WordNet para todas las categorías de los términos de la consulta TREC número 10 (expuesta en la figura 5.4) correspondiente a *"AIDS treatments"*. Asimismo, a continuación se muestra un ejemplo de expansión de la misma consulta número 10:

- Expansión total. Términos de la nueva consulta:

```
aids treatments acquired immune deficiency syndrome
aid assistance help assist assistance helping care
attention tending assist treatment discussion
discourse
```

- Expansión basada en la selección de los *synsets* mediante la desambiguación de la consulta. Términos de la nueva consulta:

```
aids treatment acquired immune deficiency syndrome
```

Como se observa, la expansión de la consulta mediante los sinónimos de WordNet consigue aumentar el número de términos de las consultas. En el ejemplo, se muestra una consulta original compuesta por dos términos y, mediante el proceso de expansión total la consulta reformulada aparece con 19 términos, mientras que con la expansión mediante la desambiguación aparece con 6.

5.2.4 Entorno de evaluación

La colección experimental de documentos y consultas

Las colecciones de prueba utilizadas en IR están formadas por un conjunto de documentos y un conjunto de consultas con sus correspondientes juicios sobre la relevancia de cada documento. Son los juicios los que permiten precisamente medir la eficacia de estos sistemas. Para nuestros experimentos hemos seleccionado la colección de documentos Wall Street Journal (WSJ), utilizada en las conferencias TREC[6] para la evaluación de sistemas de IR. En la figura 5.5 hemos incluido un ejemplo de un documento que, como puede observarse, están en formato estándar

[6] En estas conferencias participan las más importantes empresas de software en IR y muchas de las universidades que investigan en este área.

SGML, lo que facilita su tratamiento automático. Para nuestros experimentos elegimos al azar 5.000 documentos de la colección WSJ correspondientes al año 1990. También al azar se escogieron 50 consultas TREC de entre aquellas que tenían algún documento relevante en los textos seleccionados. La tabla 5.2 muestra las estadísticas, de la colección de textos —WSJ—, cuyo tamaño es 16.256 Mb. También aparecen las estadísticas referidas al número de términos que forman las consultas —TREC—.

Número de documentos	5.000
Tamaño (Mb)	16
Número de consultas	50
Número de documentos relevantes	385
Precision media en 11-pt(recall)	0,1894

Tabla 5.2: Características de la colección de prueba utilizada en la evaluación de la recuperación

Los experimentos consistirán en la evaluación de las colecciones formadas por los textos originales y un experimento (como se relata más adelante) para cada proceso de expansión de las consultas.

Dominios
Environment
Finance
International Economics
International Finance
International Relations
Law and Government
Military
Science and Technology
U.S. Economics
U.S. Politics

Tabla 5.3: Dominios de las consultas utilizadas en los experimentos

Para la evaluación de nuestros experimentos se eligirán las consultas TREC (Text REtrieval Conferences) [Sparck-Jones, 1995], formadas por las siguientes partes: dominio, título, descripción, narrativa, conceptos, factores y definiciones (ver figura 5.4). El *dominio* indica el ámbito de la consulta (en la tabla 5.3 se presentan los dominios abarcados por las consultas utilizadas en los experimentos). El *título* correspondería a una consulta formulada por un usuario de un sistema IR. La *descripción* y *narrativa* son una explicación en detalle sobre las propiedades que deben tener los documentos relevantes. La sección *conceptos* contiene una breve base de conocimiento a modo de la que podría poseer un buscador real. Los *factores* limitan algún aspecto

de la consulta, como los países sobre los que se está interesado o el momento al que se hace referencia (actual, futuro, pasado). La sección *definiciones* contiene explicaciones sobre alguno de los términos o expresiones que aparecen en las otras secciones. De toda esta información sólo utilizamos la relativa a la sección *título*, por considerarla la única representativa de una consulta típica de un usuario de un sistema de IR.

```
Noun
1. AIDS, acquired immune deficiency syndrome
-- (a serious (often fatal) disease of the immune system
transmitted through blood products especially by sexual contact or
contaminated needles)

Verb
1. aid, assistance, help
-- (a resource: "visual aids in teaching"; "economic assistance to
depressed areas")
2. aid, assist, assistance, help, helping
-- (the activity of contributing to the fulfillment of a need or
furtherance of an effort or purpose: "he gave me an assist with the
housework"; "could not walk without assistance"; "rescue party went
to their aid"; "offered his help in unloading")
3. aid -- (a gift of money to support a worthy person or cause)
4. care, attention, aid, tending
-- (the work of caring for or attending to someone or something;
"no medical care was required"; "the old car needed constant
attention")

Verb
1. help, assist, aid
-- (give help or assistance; be of service; "Everyone helped out
during the earthquake"; "Can you help me carry this table?" "She
never helps around the house")
2. help, aid
-- (improve the condition of; "These pills will help the patient")

Noun
1. treatment
-- (care by procedures or applications that are intended to relieve
illness or injury)
2. treatment, handling
-- (the management of someone or something; "the handling of
prisoners" or "the treatment of water sewage"; "the right to equal
treatment in the criminal justice system")
3. treatment
-- (a manner of dealing with something artistically; "his treatment
of space borrows from Italian architecture")
4. discussion, treatment, discourse
-- (an extended communication (often interactive) dealing with some
particular topic; "the book contains an excellent discussion of
modal logic"; "his treatment of the race question is badly biased")
```

Figura 5.3: Expansión total para los términos de la consulta número 10 (TREC) con información de WordNet

```
<num> Number: 010
<dom> Domain: Science and Technology
<title> Topic: AIDS treatments
<desc> Description:
Document will mention a specific AIDS or ARC treatment.
<narr> Narrative:
To be relevant, a document must include a reference to at least
one specific potential Acquired Immune Deficiency Syndrome (AIDS)
or AIDS Related Complex treatment.
<con> Concept(s):
1. Acquired Immune Deficiency Syndrome (AIDS), AIDS Related
Complex (ARC)
2. treatment, drug, pharmaceutical
3. test, trials, study
4. AZT, TPA
5. Genentech, Burroughs-Wellcome
<fac> Factor(s):
<def> Definition(s):
ARC - AIDS Related Complex.  A set of symptoms similar to AIDS.
AZT - Azidothymidine, a drug for the treatment of Acquired Immune
Deficiency Syndrome, its related pneumonia, and for severe AIDS
Related Complex.
TPA - Tissue Plasminogen Activator - a blood clot-dissolving drug.
treatment - any drug or procedure used to reduce the debilitating
effects of AIDS or ARC.
</top>
```

Figura 5.4: Consulta TREC (*Topic* número 10)

```
<DOC>
<DOCNO>
WSJ900416-0096
</DOCNO>
<DOCID>
900416-0096.
</DOCID>
<HL>
   International -- Washington Insight:
   Bush Is Torn Between Amity to Kaifu,
   Fear of U.S. Ire About Japanese Trade
   ----
   By Gerald F. Seib
   Staff Reporter of The Wall Street Journal
</HL>
<DATE>
04/16/90
</DATE>
<SO>
WALL STREET JOURNAL (J), PAGE A10
</SO>
<CO>
   JAPAN
</CO>
<IN>
MONETARY NEWS, FOREIGN EXCHANGE, TRADE (MON)
</IN>
<GV>
EXECUTIVE (EXE)
CONGRESS (CNG)
</GV>
<LP>
   WASHINGTON -- President Bush is in a quandary, caught
between his desire to reward a foreign friend, Japan's Prime
Minister Toshiki Kaifu, and his inclination to avoid
political firefights at home.
   The president's dilemma arises because of a politically
charged decision he must make by the end of this month:
whether to name Japan an unfair trading partner again, as he
did last year. If he does, he will set off a new round of
trade investigations and negotiations and renew the specter
of U.S. sanctions on Tokyo.
</LP>
<TEXT>
   The debate pits Bush adviser against Bush adviser. The
president earlier this year hosted Mr. Kaifu, Japan's new and
politically vulnerable leader, at a private dinner in Palm
Springs, Calif., and pressed him for basic change in Japan's
trade behavior.
      .
      .
      .
... But Mr. Bush also hears the political rumblings at home that
say it's too soon to ease the pressure. "We're not completely home
free yet," says Sen. Max Baucus of Montana, a leading Democratic
hawk on trade.
</TEXT>
</DOC>
```

Figura 5.5: Fragmento de un documento del Wall Street Journal (documento: WSJ900416-0096)

Sistema Smart de IR

Para la evaluación de nuestros experimentos hemos utilizado uno de los más conocidos sistemas experimentales de recuperación de información, el sistema SMART [Buckley, 1985], al estar basado en el modelo del espacio vectorial y dotado de un módulo de evaluación de la eficacia sobre colecciones de prueba. Asimismo, es un software de libre distribución[7] y ha sido utilizado con frecuencia por otros investigadores (como por ejemplo, Gonzalo et al. [1998a]; Hull y Grefenstette [1996]).

El sistema SMART se caracteriza por:

- Entorno del sistema:
  - Representación de vectores y cálculo de la similitud
  - Manipulación de vectores
  - Generación de vectores
- Procedimientos del sistema:
  - Indexación automática
  - Clasificación automática de documentos
  - Operaciones de realimentación por relevancia
  - Espacio de documentos dinámicos
- Mejora automática de la recuperación convencional
  - Ordenación de documentos y ponderación de términos
  - Recuperación a través de un sistema de diálogo persona-ordenador y de operaciones de *clustering* local
- Evaluación de la recuperación:
  - Cálculo de *Recall* y *Precision*

[7] Puede obtenerse a través de ftp en la Universidad de Cornell *ftp://ftp.cs.cornell.edu/pub/smart*.

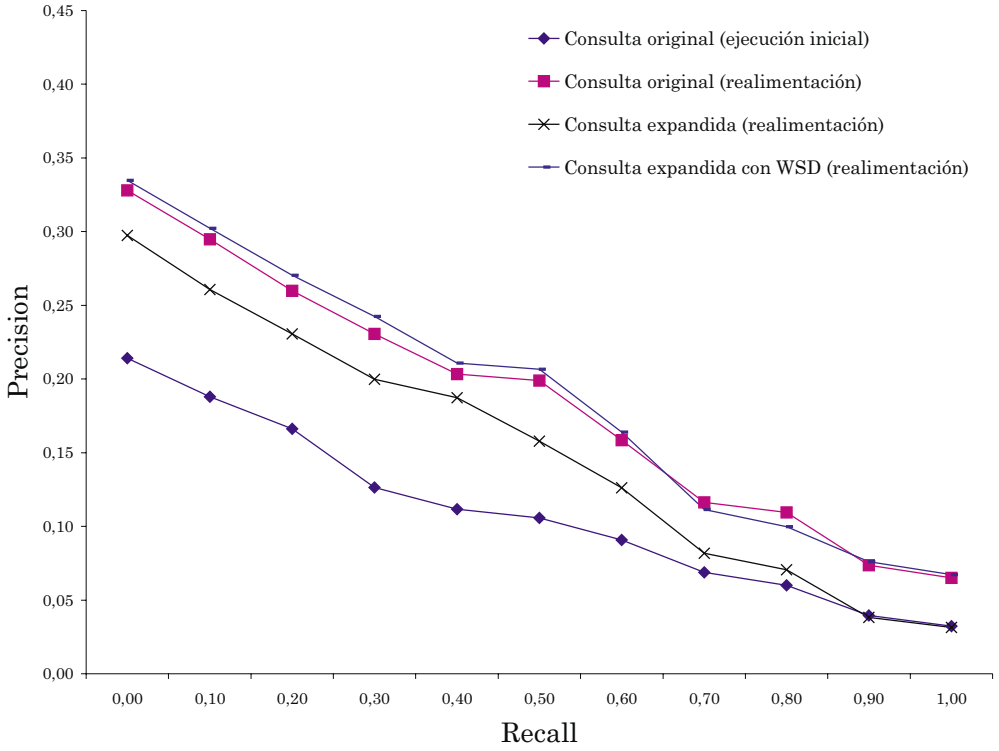

Figura 5.6: Efectividad para los diferentes tipos de expansión con *feedback*

Diseño de experimentos

Hemos realizado un conjunto de experimentos para demostrar la eficacia de la desambiguación basada en la realimentación en el proceso de expansión de consultas. Los experimentos realizados para la evaluación indirecta de la desambiguación en la recuperación, permiten comparar un enfoque de IR:

- que no realiza ningún proceso de expansión, con lo cual no incluye información proveniente de WordNet, sólo la consulta original.
- que hace uso de la expansión de consulta, mediante información de sinonimia para todos los significados y categorías gramaticales.
- que hace uso de la expansión de consulta, mediante información de sinonimia a través del sistema WSD presentado de forma automática.

Los experimentos consistieron en la evaluación de las colecciones formadas por los textos y consultas originales. Para las consultas se realizaron experimentos para cada proceso de expansión mediante información de WordNet. Los resultados que mostramos se basan en la medidas

de efectividad *recall* y *precision*, descritas en la sección 4.4 y en las fórmulas 4.1 y 4.2. La evolución pareja de *precision* y *recall* de los procesos de expansión se observa en la figura 5.6, que comentamos más adelante. Las tablas 5.4 y 5.5 contienen los datos concretos de *recall* y *precision* correspondientes a los procesos de expansión.

Resultados e interpretación de la evaluación indirecta

En la evaluación indirecta de la desambiguación que hemos realizado sobre la tarea de recuperación de información, nos interesa determinar en nuestros experimentos, hasta que punto es preciso efectuar la desambiguación cuando se expande la consulta, en nuestro caso con información de sinonimia procedente de WordNet.

	Consulta Original		Consulta Expandida	Consulta Expandida + WSD
	E. inicial	*E. feedback*	*E. feedback*	*E. feedback*
Recall				
11-pt media:	0,1094	0,1853	0,1529	0,1896
% cambio:		69,4	39,8	73,6
3-pt media:	0,1106	0,1894	0,1530	0,1922
% cambio:		71,2	38,3	73,3

Tabla 5.4: *Precision media* y porcentaje de cambio en los 11 niveles estándar y 3 intermedios (0.2, 0.5, 0.8) de *recall*

La figura 5.6 nos da una visión general muy clara del comportamiento de la expansión de consultas. Pudiendo comparar las dos expansiones realizadas con la consulta original. Los mejores valores de *precision* en el proceso de recuperación para los diferentes niveles de *recall* se obtienen en las consultas que emplean información de sinonimia con desambiguación (Consulta expandida con WSD). Produciéndose una mejora con este tipo de expansión en relación con no hacer uso de ningún recurso lingüístico (Consulta original). Por debajo aparecen las precisiones para la consulta original, seguida de la consulta expandida totalmente (para todos los significados y categorías gramaticales).

Recall	Consulta Original		Consulta Expandida	Consulta Expandida+WSD
Nivel	*E. inicial*	*E. feedback*	*E. feedback*	*E. feedback*
0,00	0,2142	0,3281	0,2975	0,3348
0,10	0,1879	0,2948	0,2607	0,3023
0,20	0,1661	0,2598	0,2306	0,2703
0,30	0,1264	0,2307	0,1998	0,2424
0,40	0,1116	0,2033	0,1873	0,2108
0,50	0,1057	0,1990	0,1579	0,2067
0,60	0,0908	0,1586	0,1262	0,1640
0,70	0,0689	0,1164	0,0818	0,1115
0,80	0,0601	0,1094	0,0705	0,0997
0,90	0,0395	0,0737	0,0382	0,0761
1,00	0,0323	0,0651	0,0315	0,0673

Tabla 5.5: Evaluación en 11 niveles de *Recall*, obtenida con realimentación por relevancia para los procesos de expansión

En la tabla 5.4 se muestra el porcentaje de cambio para el proceso de expansión en 3 y 11 niveles de *recall*. Destacando, en el caso de la expansión haciendo uso de la desambiguación (73,3%), un incremento mayor del 3,5% en 3 puntos de *recall*, con respecto a la consulta original (71,2%). Esto es un buen resultado teniendo en cuenta que se ha obtenido desambiguando automáticamente las consultas empleando la realimentación, lo que supone un incremento del 25% aproximadamente con respecto a la expansión sin desambiguación. Además, esto muestra que WordNet puede mejorar la recuperación de texto, el problema por tanto reside en la efectividad del proceso de desambiguación.

Asimismo, en la tabla 5.5 se muestran los datos con más detalle en 11 niveles de *recall*. Debe tenerse en cuenta que los valores de la *precision* en los 11 niveles de *recall* suelen ser más bajos cuando se emplea el método de la colección residual que cuando se realiza una evaluación estándar (sin hacer uso de la realimentación). La razón es que se excluyen los documentos relevantes mejor valorados. Esto no tiene gran importancia, ya que lo que realmente nos interesa medir es el incremento que se produce en la *precision* debida a la realimentación.

En la tabla 5.6 hemos recogido los incrementos que en los valores exactos de *precisión* y *recall* se obtienen con la expansión de consultas. Como se observa en ella, los incrementos de *recall* relativos a la expansión de consulta con WSD son importantes (aproximadamente del 62%) comparados con la consulta original (*e. inicial*). Asimismo, los incrementos en *precision* son del orden del 28% mayor que los obtenidos por la consulta original (*e. inicial*).

La utilización de sinónimos permite recuperar otros documentos relevantes que no contenían las mismas palabras que las utilizadas por el usuario en su consulta, lo que provoca un aumento del

	Consulta Original		Consulta Expandida	Consulta Expandida + WSD
	E. inicial	*E. feedback*	*E. feedback*	*E. feedback*
Recall Exact:	0,1782	0,2838	0,2242	0,2901
en 5 docs	0,0753	0,1377	0,0970	0,1364
en 10 docs	0,1328	0,2299	0,1693	0,2419
en 15 docs	0,1782	0,2838	0,2242	0,2901
en 30 docs	0,2837	0,3629	0,3106	0,4251
Precision Exact:	0,0711	0,0978	0,0957	0,0913
en 5 docs	0,0933	0,1422	0,1348	0,1261
en 10 docs	0,0778	0,1178	0,1130	0,1130
en 15 docs	0,0711	0,0978	0,0957	0,0913
en 30 docs	0,0578	0,0711	0,0688	0,0717

Tabla 5.6: *Precision* y *Recall* en 5, 10, 15 y 30 documentos

recall. Por el contrario, también se seleccionan algunos otros por contener "supuestos sinónimos", por ejemplo, palabras que WordNet proporciona como sinónimos pero que se emplean en la consulta con significado diferente al que utiliza el usuario. Esta situación provoca una pérdida de *precision*, como es el caso de la expansión total.

Una interpretación que hacemos de los datos obtenidos, es que WordNet presenta atributos positivos, pero también limitaciones. En concreto, incluye demasiada granularidad en la distinción de sentidos y no dispone de dominios de información para dicha tarea. Además, varios autores Gonzalo et al. [1998b] argumentan que los procesos de expansión son dependientes de los recursos utilizados en la tarea de recuperación de texto.

	Consulta Original	Consulta Expandida+WSD
Recall exact:	0,0570	0,0730
Precision exact:	0,0166	0,0176

Tabla 5.7: Valores exactos de *precision* y *recall* para consultas con 1 y 2 términos

Por otra parte, un aspecto que nos interesa destacar, es el aumento significativo de la efectividad, cuando las consultas son cortas o incompletas. En la tabla 5.7 se muestran valores medios de *precision* y *recall* para 10 consultas cortas elegidas al azar de entre las que tienen uno y dos términos. Los valores exactos se exponen, tanto para la Consulta Original, formada por 10 consultas sin expandir, como para la Consulta Expandida con WSD, en la que se han expandido las 10 consultas originales.

En experimentos previos de trabajos relacionados, nos resulta especialmente relevante para

	precision	% cambio
Consultas no-expandidas	0,1634	—
N=70.000;		
α=0,3	0,1627	-0,5%
α=0,5	0,1603	-1,9%
α=0,8	0,1543	-5,6%
N=35.000;		
α=0,3	0,1636	0,1%
α=0,5	0,1635	0,1%
α=0,8	0,1639	0,3%

Tabla 5.8: Efectividad de la estrategia de Voorhees [1994] cuando se expande la consulta con los *synsets* seleccionados automáticamente

la interpretación de nuestros resultados los obtenidos por Voorhees [1994]. En la tabla 5.8 se muestra una síntesis de los valores obtenidos experimentalmente en el proceso de expansión de consultas realizado por Voorhees. En concreto se expone la *precision* media en 11 puntos de *recall* y el porcentaje de cambio entre la consulta original (sin expandir) y la consulta expandida con información de sinonimia. El proceso de expansión realiza automáticamente la selección de *synsets*. Se puede observar la disminución de la efectividad del proceso de recuperación con expansión mediante la selección automática de *synsets*. A pesar de que los resultados obtenidos no son comparables de forma directa con los nuestros, apreciamos que la estrategia de Voorhees no mejora la efectividad de la recuperación. Sin embargo, en nuestro caso, aunque en general no conseguimos una mejora significativa, en particular para el caso de consultas cortas, sí se obtiene una mejora en el proceso de recuperación.

Concluyendo, un aspecto importante son los pobres resultados de la expansión de consulta para todos los sinónimos (Consulta expandida), ya que ésta no mejora los resultados de la consulta original, sino que empeoran como consecuencia de la introducción de muchos términos con diferentes significados. Sin embargo, mediante la resolución de la ambigüedad, como se muestra en la última columna de la tabla 5.4 (Consulta Expandida con WSD), la *precision* aumenta significativamente con respecto a la Consulta Expandida y ligeramente en relación con la Consulta Original. Esto se debe fundamentalmente, a que el proceso de expansión se ha realizado selectivamente y de manera automática por medio de la desambiguación.

Por otra parte, con nuestros resultados se pone de manifiesto que el empleo de la realimentación proporciona información contextual a las consultas y permite solventar el problema de resolución automática de la ambigüedad léxica, citado entre otros, por Voorhees [1993, 1994]. Finalmente, comentar de acuerdo con Voorhees, que la expansión no mejora significativamente la efectividad cuando los consultas son relativamente completas, sin embargo las consultas más cortas, menos elaboradas, mediante la expansión pueden mejorar la efectividad de la recuperación.

5.3 Integrando recursos lingüísticos en TC por medio de WSD

La categorización de textos es una tarea muy importante en el marco del acceso a la información. Como se ha adelantado en el Capítulo 2, consiste en la clasificación de documentos dentro de un conjunto de categorías predefinido. La categorización automática de textos es una tarea compleja de clasificación, frecuentemente aplicada a la asignación de descriptores de contenido a documentos, al encaminamiento y filtrado de texto, o empleada como parte de otros sistemas de procesamiento del lenguaje natural [Lewis, 1992].

Los enfoques más habituales de categorización de documentos se basan en la utilización de una colección de documentos previamente etiquetados (colección de entrenamiento) para predecir la asignación de categorías a nuevos documentos[8]. Suponiendo que los términos que aparecen con frecuencia en presencia de una determinada categoría y con menos frecuencia en presencia de otras, son buenos indicadores de aparición de esa categoría, como se muestra en la figura 5.7. Las frecuencias de aparición se calculan sobre un corpus de documentos categorizado. La idea general es obtener una representación de cada categoría a partir de los documentos de entrenamiento, con la que se compara la representación de cada nuevo documento y se decide si se incluye este último en la categoría. Sin embargo, es corriente la existencia de categorías con pocos documentos de entrenamiento, lo que dificulta la obtención de representaciones eficaces de las mismas.

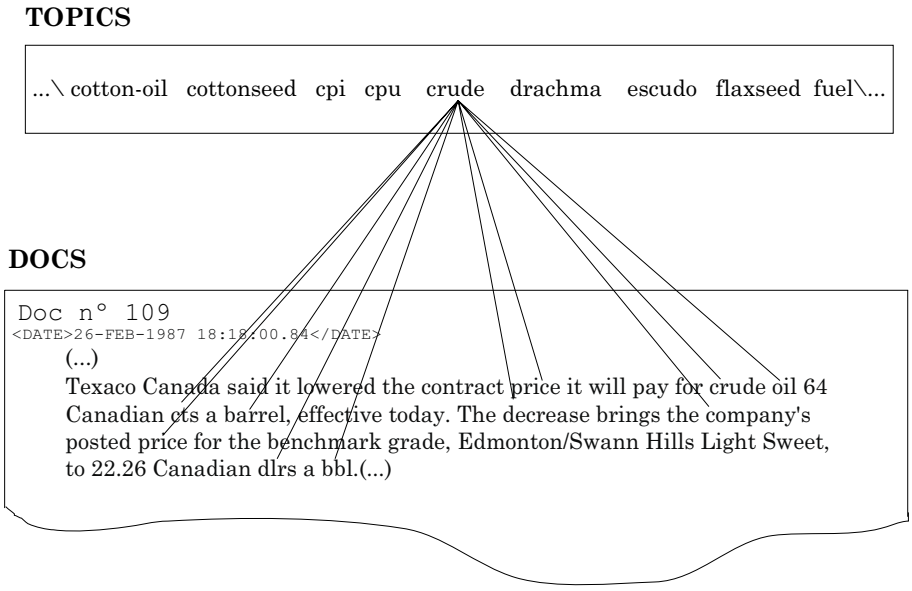

Figura 5.7: Términos que aparecen cerca de una categoría

La utilización de recursos lingüísticos adicionales puede mejorar la representación de las categorías, y en consecuencia, la eficacia de los sistemas de categorización [Buenaga et al., 1997;

[8] En [Yang, 1999; Yang y Liu, 1999; Sahami, 1998] se ofrecen comparativas.

Díaz et al., 1998; Ureña et al., 2001]. En este estudio hemos tomado como base un método de categorización de textos basado en la integración de recursos lingüísticos [Buenaga et al., 1997].

Las bases de datos léxicas son un recurso valioso, ya que pueden proporcionar información sobre los términos, y así mejorar la representación de las categorías. De manera análoga al sistema de desambiguación presentado en el Capítulo 4, se puede utilizar la base de datos léxica WordNet [Miller, 1995] como recurso complementario a la colección de entrenamiento. Para la utilización de WordNet de manera eficaz, es preciso efectuar un proceso de desambiguación de las categorías que aparecen en la colección de documentos, como se detalla en las dos próximas secciones.

Los documentos y las categorías se representan en términos de vectores de pesos, donde cada componente representa la importancia de un término en un documento o categoría. Los pesos para los documentos se computan con la clásica fórmula $tf \cdot idf$[9] [Salton y McGill, 1983].

$$ws_{ij} = t_{ij} \cdot log_2(n/f_i) \tag{5.4}$$

Donde t_{ij} es la frecuencia del término i en el documento j, n es el número total de documentos, y donde f_i el número de documentos donde el término i aparece. Después, se calculan los pesos wc_{ik} para los vectores de categorías, para posteriormente calcular la similitud entre los documentos de prueba ($\vec{d_j}$) y las categorías ($\vec{c_k}$) de manera uniforme a la realizada en la fórmula 4.11 para la desambiguación:

$$sim(\vec{d_j}, \vec{c_k}) = \frac{\sum_{i=1}^{N} ws_{ij} \cdot wc_{ik}}{\sqrt{\sum_{i=1}^{N} ws_{ij}^2 \cdot \sum_{i=1}^{N} wc_{ik}^2}} \tag{5.5}$$

5.3.1 Uso de WordNet en la categorización

La idea que se pretende, es extraer de WordNet, información para definir el contexto semántico de una categoría. WordNet proporciona entre otras relaciones, como se ha comentado en el Capítulo 3, la de sinonimia. Aunque se puedan establecer complejas relaciones usamos solamente la información de sinonimia, esto es, aprovechamos WordNet como thesaurus y así complementar la información de la colección de entrenamiento[10]. La utilización de otras relaciones como la generalización se ha explorado con resultados negativos en otros trabajos [Scott y Matwin, 1998].

Hemos de definir conjuntos de sinónimos a partir de las palabras que constituyen las categorías. La técnica más simple es usar los *synsets* de WordNet como conjuntos de sinónimos.

[9] TF (Term Frequency) y IDF (Inverse Document Frequency).

[10] Básicamente, el proceso de entrenamiento asigna pesos a los términos de la colección Reuters, en vectores de categorías; en proporción al número de apariciones del término en la categoría, y en proporción a la importancia del término en la colección.

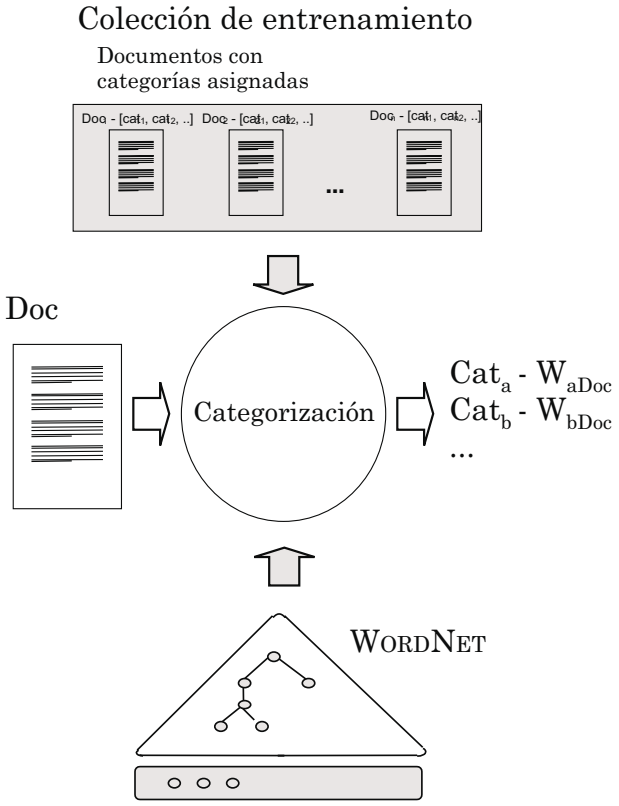

Figura 5.8: Proceso de integración de recursos lingüísticos

La combinación de información de WordNet y de la colección de entrenamiento se realiza por el uso inicial de pesos para clases (ver figura 5.8[11]). La utilización de WordNet se basa en la asunción de que los nombres de las categorías pueden ayudar a predecir sus ocurrencias [Buenaga et al., 1997].

Por ejemplo, la ocurrencia de la palabra "barley" sugiere que nuevos artículos podrían clasificarse dentro de la categoría *barley* de Reuters-21578.

Así, los nombres de las categorías se buscan en WordNet, para obtener un conjunto de sinónimos que represente a cada categoría. Se construye un vector inicial de pesos para cada categoría, donde cada componente es el peso del término en la colección.

Por ejemplo, si buscamos en WordNet el nombre para la categoría *earnings*[12], obtenemos dos sentidos o *synsets*:

```
1. Net income, net, net profit, lucre, profit, profits, earnings
--(the excess of revenues over outlays in a given period of time)
```

[11] Extraída de [Vaquero y Buenaga, 1996].

[12] Los ejemplos presentados corresponden a palabras reales y categorías de Reuters-21578 y WordNet 1.6.

```
2. Wage, pay, earnings, remuneration, salary --(something that
remunerates; "wages were paid by check"; "he wasted his pay on
drink"; "they saved a quarter of all their earnings")
```

De manera general esta propiedad se utiliza como sigue:

- Se busca el código o expresión asociada a cada categoría en WordNet, obteniéndose una lista de *synsets* asociados a la misma.
- Se selecciona el *synset* más adecuado para representar la categoría.

De esta forma, los términos se obtienen de los *synsets* seleccionados, después de realizar un proceso de filtrado mediante un algoritmo de listas de parada *stoplists*. Para cualquier término, el grado de proximidad semántica a una categoría se calcula a través de los siguientes criterios:

- Si el término es un sinónimo directo de la expresión que representa la categoría (por ejemplo, el término "profit" es un sinónimo de la expresión "earnings", que corresponde a la categoría con el mismo nombre "earnings") , la proximidad semántica entre el término y la categoría se establece a 1.
- Si la expresión que representa a una categoría consta de varios términos, y de manera conjunta la expresión no se encuentra en WordNet, se buscan los términos individualmente que componen la expresión nuevamente en WordNet[13]. En esta situación el valor semántico para los sinónimos asociados a estos términos se define como $1/n_c$, siendo n_c el número de palabras en la expresión.
- Si se pueden definir varios valores entre una categoría y un término, se selecciona el valor mayor.

Para las 135 categorías en la colección de documentos Reuters, se obtuvieron un conjunto de 246 términos. También se generaron un conjunto de 346 valores de la proximidad semántica término-categoría. Éstos fueron tomados como una representación inicial de las categorías. Se ha obtenido un peso w_i para cada uno de los términos de la colección de entrenaniento, mediante la fórmula *idf* de Salton y McGill [1983]:

$$w_i = log_2(n/f_i) \tag{5.6}$$

La selección de los *synsets* con mayor proximidad semántica a la categoría es necesaria, ya que la inclusión de demasiados términos podría conducir a un categorizador de texto poco efectivo. Esto es un problema de ambigüedad léxica.

El proceso de selección, en definitiva, es un proceso de desambiguación, que hemos realizado manual y automáticamente a través de los métodos presentados en el capítulo anterior.

[13]Por ejemplo, el término "indicant" es un sinónimo de la palabra "index" en la expresión "industrial production index" (correspondiente a la categoría con código IPI), el valor de proximidad semántica es 1/3.

5.3.2 Uso de la desambiguación para la categorización basada en WordNet

Siguiendo con el método para la integración de recursos lingüísticos en la categorización de textos, consideramos que un aspecto importante es la desambiguación de términos. En este marco introducimos nuestro método de resolución de la ambigüedad léxica.

La selección del *synset* más adecuado para representar una categoría, como se ha visto en la sección anterior, es en definitiva un proceso de desambiguación.

Intuitivamente, cada categoría tiene un solo significado en la colección de documentos, y los experimentos realizados, demuestran que es preciso realizar la desambiguación para sacar el máximo partido a la información de WordNet.

La desambiguación puede realizarse de manera manual, pero es conveniente mecanizar este proceso a fin de obtener un sistema de categorización completamente automático.

El proceso de desambiguación de cada una de las categorías se ha realizado haciendo uso del sistema WSD propuesto en el Capítulo 4, particularizado para esta tarea TC, como sigue:

- Para cada una de las categorías (*Topics*) se obtienen todos los *synsets* (conjunto de sinónimos). Para cada *synset* se calcula su ventana contextual mediante la construcción de un vector de pesos, donde el vector para la categoría i con sentido j es $s_{ji} = \langle ws_{j1}, ws_{k1}, ..., ws_{kn} \rangle$. Donde ws_{ki} son los pesos de las palabras circundantes. Esta información se complementa con SemCor de manera similar.

- De forma análoga, para cada una de las categorías a ser desambiguadas se construye la ventana contextual con la información que proporciona el corpus de entrenamiento Reuters. Esto representa la consulta para el desambiguador. El vector de pesos para c_k, es $c_k = \langle wc_1, wc_{k1}, ..., wc_{kn} \rangle$.

- Se calcula la similitud de la palabra (categoría) para cada uno de los sentidos posibles. Haciendo uso de las expresiones 4.11 y 4.14, descritas en desambiguación.

5.3.3 Experimentos centrados en la efectividad

Los experimentos realizados para la evaluación indirecta de la desambiguación en la categorización, permiten comparar un enfoque de TC:

- que no incluye información proveniente de WordNet.

- que hace uso de WordNet, a través del sistema WSD presentado de forma automática.

- que hace uso de WordNet, mediante un perfecto método de desambiguación, manual que es el que representa las decisiones humanas.

Estos enfoque se comparan con un algoritmo *línea base*, donde el proceso de categorización de texto no hace uso de información procedente de WordNet, es decir, se basa sólo en el entrenamiento.

		Subcolección		
		Entrenamiento	Prueba	Total
Docs.	Número	13.625	6.188	19.813
Palabras	Número	1.820.881	746.726	2.567.607
	Media	133	120	129
Docs. con 1+Topics	Número	7.780	3.022	10.802
	Porcentaje	57	48	54
Topics	Número	9.666	3.768	13.434
	Media	0,70	0,60	0,67

Tabla 5.9: Estadísticas de la colección de documentos Reuters-21578

Recursos empleados en los experimentos

Para los experimentos de categorización presentados en esta memoria, se han utilizado la colección de evaluación Reuters-21578 y la base de datos léxica WordNet. La colección Reuters ha sido la más utilizada en la evaluación de sistemas de categorización, y permite, la comparación con otros enfoques [Yang, 1999].

Para la evaluación de la categorización, se suele dividir la colección en dos partes, una de entrenamiento y otra de prueba. Nosotros hemos usado la partición de Lewis [Lewis, 1992], una de las más populares, que reserva dos terceras partes de los documentos para entrenamiento y una tercera parte para evaluación. Podemos ver en la tabla 5.9 una estadística de la partición, referente al tamaño de los documentos de la colección Reuters y el número de documentos con uno o más *Topics* asignados, así como el número de *Topics* asignados a los documentos. En esta partición existen 89 categorías que poseen documentos en las subcolecciones de entrenamiento y de evaluación.

Existen varias medidas utilizadas para evaluar la efectividad de la categorización, que se pueden clasificar en medidas para clasificadores basados en *ranking*, y en medidas para clasificadores binarios [Yang, 1999]. Nuestro categorizador produce un *ranking* de documentos para cada categoría, que se puede convertir en una asignación binaria (el documento se introduce en la categoría o no). Nosotros presentamos los resultados, utilizando tanto medidas basadas en *ranking* (*precision* a 11 niveles de *recall*) como medidas para categorizadores binarios (medida *f1* calculada realizando *macroaveraging* y *microaveraging*) [Yang, 1999].

Resultados e interpretación de la evaluación indirecta

El propósito de la evaluación indirecta es determinar si es preciso efectuar la desambiguación de las categorías. La evaluación se realiza sobre las 85 categorías que tienen apariciones en SemCor, además de poseer documentos en la subcolección de entrenamiento y en la de evaluación de la partición Lewis de Reuters. En el Apéndice D se incluye una relación de las categorías utilizadas en la desambiguación. Además se muestra la lista completa de las categorías del conjunto *Topics*,

	TC		**TC+WSD**		
Recall	*Training*	*Training+WN*	*Rocchio*	*Widrow-Hoff*	*hand*
0,0	0,805	0,880	0,900	0,900	0,889
0,1	0,777	0,851	0,881	0,881	0,872
0,2	0,733	0,822	0,857	0,857	0,844
0,3	0,683	0,764	0,818	0,818	0,808
0,4	0,623	0,708	0,770	0,770	0,778
0,5	0,570	0,658	0,731	0,731	0,744
0,6	0,500	0,580	0,608	0,608	0,624
0,7	0,404	0,496	0,532	0,532	0,550
0,8	0,337	0,415	0,466	0,466	0,474
0,9	0,256	0,317	0,346	0,346	0,360
1,0	0,124	0,190	0,217	0,217	0,227
Average	0,528	0,608	0,648	0,648	0,652

Tabla 5.10: *Precision* en 11 niveles de *recall*

junto con la expresión a la que hacen referencia (por ejemplo, L-cattle hace referencia a "live cattle"), y su significado en español.

En las tablas 5.10 y 5.11 se presentan los resultados de la evaluación de la categorización realizada solo con entrenamiento sin WordNet (*Training*), utilizando todos los *synsets* por categoría (*Training+WordNet*), desambiguando con el algoritmo de Rocchio (*Rocchio*), desambiguando con el algoritmo de Widrow-Hoff (*Widrow-Hoff*), y desambiguando manualmente (*hand*). En la tabla 5.10 se muestran los resultados de la categorización utilizando la medida de *precision* en 11 niveles de *Recall* y *Precision* media (*Average*). Asimismo, la figura 5.9 nos da una visión general muy clara del comportamiento de los diferentes tipos de categorización. Los valores mejores de *precision* para los diferentes niveles de *recall* los obtienen los categorizadores o procesos que hacen uso de WordNet y de la desambiguación. Se observa claramente que la *precision* alcanzada por el desambiguador automático coincide prácticamente con el desambiguador humano (manual). Como estábamos apuntando, no existe una gran diferencia entre los resultados obtenidos haciendo uso de la desambiguación automática en comparación con la manual.

	TC		**TC+WSD**		
f_1	*Training*	*Training+WN*	*Rocch.*	*Widrow.*	*hand*
Macroaveraging	0,464	0,538	0,571	0,571	0,576
Microaveraging	0,661	0,664	0,674	0,674	0,678

Tabla 5.11: f_1 calculada por medio de *macro* y *microaveraging*.

En la tabla 5.11 se muestran los resultados de la categorización medidos, utilizando la *f1* calculada tanto efectuando *macroaveraging* como *microaveraging*. En cada tabla se compara la

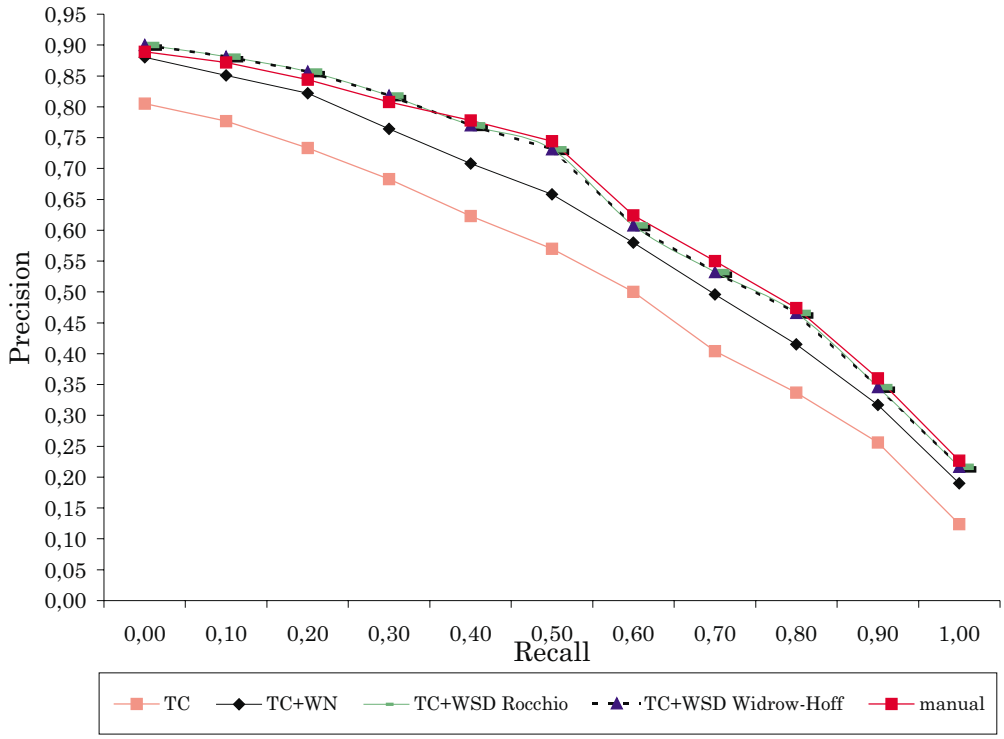

Figura 5.9: Efectividad para los diferentes tipos de categorización

categorización sin WordNet, la categorización utilizando todos los *synsets* para cada categoría, la categorización con desambiguación usando Rocchio, la categorización con desambiguación usando Widrow-Hoff, y la categorización con desambiguación manual. Se pueden realizar las siguientes observaciones:

- Globalmente, los resultados apoyan la tesis de que para sacar máximo partido de WordNet en el enfoque de categorización propuesto, es conveniente utilizar algún método de desambiguación. Si no lo hay disponible, también da buenos resultados utilizar todos los *synsets* asociados a cada categoría.

- Los dos métodos empleados en la desambiguación producen exactamente los mismos resultados. Esto se produce porque, dado el pequeño número de categorías, ambos algoritmos presentan pocas discrepancias en la elección de los *synsets* de WordNet. Como además los *synsets* potenciales poseen muchas palabras en común, no existe ninguna discrepancia en la elección de los términos para representar a las categorías entre ambos algoritmos. Por ejemplo, la categoría "inventories" posee 5 significados como se muestra en la figura 5.10. El significado elegido por el algoritmo basado en Widrow-Hoff es el segundo que es correcto. Sin embargo, aunque Rocchio elige el primer significado, los términos aportados por

ambos enfoques son iguales salvo por "list", que aparece unas 780 veces en la colección de entrenamiento, y tiene un peso tan pequeño que no afecta a la eficacia de la categorización.

```
The noun "inventories" has 5 senses (first 4 from tagged texts)

 1. inventory, stock list --
    (a detailed list of all the items in stock)
 2. stock, inventory -- (the merchandise that a shop has on hand;
    "they carried a vast inventory of hardware")
 3. inventory --
    ((accounting) the value of a firm's current assets including
    raw materials and work in progress and finished goods)
 4. armory, armoury, inventory --
    (a collection of resources; "he dipped into his intellectual
    armoryto find an answer")
 5. inventory, inventorying, stocktaking --
    (making an itemized list of merchandise or supplies on hand;
    "the inventory took two days")
```

Figura 5.10: Consulta en WordNet de la categoría "inventories"

Es interesante observar que las diferencias considerando f_1 calculada utilizando *microaveraging* son apenas apreciables. Dado que las medidas utilizando *microaveraging* dan más peso a las categorías con más documentos, esto demuestra que las mayores mejoras se producen precisamente para las categorías con menos documentos (de entrenamiento).

Finalmente, señalar que los trabajos realizados en clasificación de texto presentan los siguientes inconvenientes:

- Sólo se comparan los resultados de la clasificación usando desambiguación manual y sin usar ningún tipo de desambiguación. No se estudian métodos prácticos de desambiguación.
- Las mejoras que potencialmente debe producir la desambiguación en la clasificación son muy dependientes del modo en que la primera se utilice al servicio de la última.
- Algunas tareas pueden ser más sensibles que otras a los errores producidos en la desambiguación.

El modo en que se aplica la desambiguación permite obtener mejoras substanciales en la categorización.

5.4 Resumen y conclusiones

En este capítulo hemos estudiado la aplicación de la resolución de la ambigüedad léxica en el proceso de dos tareas de clasificación automática de documentos, en concreto, en la recuperación de información y categorización de texto.

En la recuperación de información utilizamos la técnica de realimentación para adquirir información contextual y poder resolver la ambigüedad de los términos de la consulta. En la desambiguación hacemos uso de un enfoque basado en la combinación de recursos lingüísticos para mejorar la efectividad de los sistemas de recuperación de información. A pesar de la complejidad de la tarea y la falta de información contextual de algunas consultas, se ha mejorado ligeramente la efectividad de la recuperación haciendo una expansión de los términos de la consulta, una vez desambiguada ésta empleando la realimentación, mediante información proveniente de WordNet. Los resultados demuestran que es preciso hacer uso de la desambiguación para sacar el máximo partido a la expansión de la consulta con WordNet.

También hemos empleado la resolución de la ambigüedad léxica para el caso de la categorización de texto, en concreto, en un enfoque basado en la integración de varios recursos lingüísticos[14]. Esta integración se ha realizado automáticamente a través de la desambiguación. Los resultados de los experimentos muestran que la integración de recursos en general es un enfoque muy efectivo para tareas de análisis del contenido textual. Asimismo, que la desambiguación es necesaria para realizar efectivamente el proceso automático de integración de recursos lingüísticos (de utilización de WordNet) en TC, y así obtener un beneficio efectivo.

También hemos expuesto un método de evaluación automático para ambas tareas que nos permite comparar la efectividad de la recuperación de textos y categorización.

[14]Presentando un marco uniforme para tareas de análisis de textos, basado en la integración de recursos lingüísticos.

Capítulo 6

Conclusiones

En este capítulo presentamos una descripción de las principales contribuciones presentadas en esta tesis y concluimos exponiendo los futuros desarrollos.

6.1 Principales aportaciones

Primeramente, hemos descrito las tareas de análisis del contenido y estudiado la resolución de la ambigüedad léxica y clasificación de documentos estableciendo paralelismos y elementos próximos entre ambos campos. Así, hemos introducido una serie de elementos para la desambiguación, utilizados dentro del proceso de recuperación de información, que constituyen la base para la implementación de nuestro sistema de resolución de la ambigüedad léxica. Seguidamente, hemos analizado y expuesto los recursos lingüísticos disponibles más relevantes para nuestro modelo y, los distintos métodos de desambiguación.

A continuación, resumimos las principales contribuciones de esta investigación:

1. Se ha propuesto un nuevo enfoque para la resolución de la ambigüedad léxica basado en la integración de recursos lingüísticos, para ello se utiliza información:

   - proveniente de un corpus de textos, SemCor, y
   - de una base de datos léxica, WordNet. Cuya información viene dada por las relaciones léxicas y semánticas de dicha base de datos.

2. Evaluación directa de la desambiguación: hemos mostrado de forma experimental, sobre un amplio conjunto (colección de prueba), la efectividad de nuestra aproximación de desambiguación de términos basada en la integración de recursos lingüísticos empleando una evaluación automática. A través de experimentos comparativos confirmamos que el enfoque utilizando varios recursos obtiene mejores resultados que empleando los recursos aisladamente.

3. Se ha aplicado la resolución de la ambigüedad léxica a dos tareas concretas de clasificación de documentos: categorización de textos y recuperación de información.

   - En el proceso de recuperación de información se ha aplicado la desambiguación basada en la realimentación. A pesar de la complejidad de la tarea, se ha mejorado la efectividad de la recuperación haciendo una expansión de los términos de la consulta, una vez desambiguada ésta empleando la realimentación, mediante información de sinonimia de WordNet. Los resultados demuestran, por una parte, que la técnica de realimentación permite obtener información contextual para desambiguar cuando el número de términos de la consultas es pequeño. Por otra parte, que es preciso hacer uso de la desambiguación para sacar el máximo partido a la expansión de la consulta con WordNet. Para la evaluación se han utilizado una colección de documentos (WSJ) y otra de consultas (TREC).
   - En la categorización de textos se ha propuesto la resolución automática de la ambigüedad léxica en un enfoque también basado en la integración de recursos lingüísticos. El enfoque de categorización está basado en la integración del corpus Reuters y la base de datos léxica WordNet. Este es un enfoque novedoso al incorporar la desambiguación automática en el proceso de integración de recursos lingüísticos en la tarea de categorización de textos.
   - Evaluación indirecta de la desambiguación: hemos expuesto y evaluado ambas tareas mediante un método sistemático que nos ha permitido comparar la efectividad en el ámbito de los sistemas de clasificación de documentos, tanto en recuperación de información como en categorización de textos.

6.2 Futuros desarrollos

Como principal línea de trabajo consideramos el estudio sistemático de la sensibilidad de las tareas de análisis del contenido a los errores de la desambiguación, en concreto profundizar en la categorización de textos y recuperación de información. En esta última estudiaremos también la sensibilidad con otras otras colecciones.

Plantearemos el reto, de desarrollar nuevos experimentos de aplicación de la resolución de la ambigüedad léxica a otras tareas específicas del procesamiento del lenguaje natural, en la cual la desambiguación pueda se útil. Especialmente trataremos la recuperación de información multilingüe.

Por otra parte, valoraremos la aportación que puede realizar la desambiguación, junto con la generación automática de resúmenes aplicadas al mismo tiempo a un sistema de recuperación de información que incorpore realimentación. El resumen del documento generado podría utilizarse para realimentar el sistema y también, como contexto suficiente para desambiguar el sentido de las palabras de la consulta. De esta manera podríamos expandir la consulta incorporando información procedente de WordNet y de los resúmenes valorados por el usuario.

Otra línea en la que podemos dirigir nuestro trabajo, es la de proponer un marco para la evaluación indirecta de distintos métodos propuestos de desambiguación, ya que los beneficios de la desambiguación se deben traducir en aplicaciones del mundo real.

Finalmente, esperamos que nuestro método pueda facilitar y reducir el trabajo humano en el proceso de anotación semántica de corpora de textos, al ser éstos muy necesarios para diferentes aplicaciones.

Apéndice A

SemCor

A.1 Estadísticas

	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 semánticos a adjetivos satélites.	9.970	8.347	0	18.317
punteros a nombres propios	5.602	4.075	7	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
sentidos apuntados por adjetivos satélites.	3.451	3.051	0	6.502

Tabla A.1: Estadísticas de SemCor

A.2 Estructura de los documentos

```
CONTEXTFILE ::= <contextfile concordance= conc >
       CONTEXT+
       </contextfile>
CONTEXT ::= <context filename= filename paras=yes>
       PARA+ | SENT+
       </context>
PARA ::= <p pnum= paragraph_number >
       SENT+
       </p>
SENT ::= <s snum= sentence_number >
       SENT_TOK+
       </s>
SENT_TOK ::= ( WORD_FORM | PUNC )+
WORD_FORM ::= <wf cmd=tag RDF SEP POS > word </wf>
       | <wf cmd=ignore DC SEP POS > word </wf>
       | <wf cmd=done RDF SEP POS SEM_TAG OT> word </wf>
       | <wf cmd=(update | retag) RDF SEP POS TAGNOTE NOTE>
       word </wf>
POS ::= pos= POS_TAG
POS_TAG ::= CC | CD | DT | EX | FW | IN | JJ | JJR | JJS |
           LS | MD | MD | VB | NN | NNP | NNPS| NNP| NP |
           NNP | VBN | NNS | NN|SYM | NP | NPS | PDT |
           POS | PP | PR | PRP | PRP$ | RB | RBR | RBS |
           RP | TO | UH | VB | VBD | VBG| VBN | VBP | VBZ |
           WDT | WP| WP$ | WRB
SEM_TAG ::= LEMMA WNSN LEXSN PN | NULL LEMMA ::= lemma= lemma
WNSN::= wnsn= sense_number LEXSN ::= lexsn= lex_sense PN ::=
pn=CATEGORY | NULL CATEGORY ::= person | location | group |
other RDF ::= rdf= redefinition | NULL DC ::= dc= distance |
NULL SEP ::= sep=" separator_string " | NULL TAGNOTE ::=
tagnote= TAGNOTE_TYPE TAGNOTE_TYPE ::= sns_miss | indist_sns |
wd_miss | insuffctxt | sense_lost | misc NOTE ::= note=
" note " OT ::= ot= OTHER_TAG | NULL OTHER_TAG ::= notag |
metaphor | idiom | complexprep | foreignword | nonceword
PUNC ::= <punc>
PUNC_CHARACTER</punc> PUNC_CHARACTER ::= [ , . ? ! , ; ( [ )
 ] ‘’$ " : ]
```

A.3 Descripción de los elementos SGML

cmd	Significado
tag	palabra etiquetada
done	palabra etiquetada semánticamente
ignore	palabra no etiquetada

Tabla A.2: Valores de `cm`

`<contextfile concordance=conc>`

Este elemento indica el comienzo de un fichero de contexto, especificando el nombre de la concordancia semántica que se encuentra en el fichero.

`<context filename=filename paras=yes>`

Este elemento indica el comienzo de un documento, `filename` es el nombre del fichero del corpus original que se ha extraído. `paras` indica que este documento contiene delimitadores de párrafos.

`<p pnum=paragraph-number>`

Comienzo de un nuevo párrafo. `paragraph-number` es un entero. El primer párrafo en un contexto se numera con 1, y los números de los párrafos son incrementados secuencialmente.

`<s snum=sentence-number >`

Comienzo de una nueva frase. `sentence-number` es un entero. La primera frase de cada documento se numera con 1, y los números de frases se incrementan secuencialmente en todo el documentos. El número de frases no vuelve a comenzar en 1 en cada párrafo.

`<wf attribute/value-pairs > word </wf>`

Este elemento representa un formulario de palabras. `word` es la forma de representación tal y como aparece en el documento original. Todo acerca de la información sintáctica y semántica queda almacenado en la pareja `attribute/value-pairs` de acuerdo con:

`cmd= cmd` Indica el estado del elemento formulario de las palabras (wf) (Tabla A.2).

`pos=pos pos` etiqueta sintáctica asignada. En la tabla A.3 se pueden ver las etiquetas sintácticas con sus posibles valores.

`lemma=lemma` Es el lema del término.

`wnsn=sense-number sense-number` es un entero que representa el número de sentido correspondiente al sentido de WordNet.

`pn=category` Indica que el término es un nombre propio.

`rdf=redefinition` Si esta presente, es que la palabra está redefinida con algo más.

`dc=distance` Indica que el término es una parte de una colocación discontinua en la que los términos que comprenden la colocación no son adyacentes. `distance` es un entero que especifica cuantos elementos wf permanecen fuera para la etiqueta semántica de esa colocación.

`ot=other-tag` Si esta presente, no se puede asignar una etiqueta semántica.

A.4 Etiquetas sintácticas

Etiqueta sintáctica	Interpretación
CC	Conjunción coordinada
CD	Número cardinal
DT	Determinante
EX	Existencial "there"
FW	Palabra desconocida
IN	Preposición o conjunción subordinada
JJ	Adjetivo
JJR	Adjetivo, comparativo
JJS	Adjetivo, superlativo
LS	Marcador de items
MD	Modal
NN	Nombre en singular
NNP	Nombre propio en singular
NNPS	Nombre propio en plural
NNS	Nombre en plural
NP	Nombre propio en singular
NPS	Nombre propio en plural
PDT	Predeterminante
POS	Terminación de un posesivo
PP	Pronombre personal
PR	Pronombre
PRP	Pronombre
PRP$	Pronombre plural
RB	Adverbio
RBR	Adverbio de comparación
RBS	Adverbio superlativo
RP	Participio
SYM	Símbolo
TO	to
UH	Interjección
VB	Forma base de un verbo
VBD	Pasado de un verbo
VBG	Gerundio o participio pasado de un verbo
VBN	Participio pasado de un verbo
VBP	Verbo no en tercera persona del presente singular
VBZ	Verbo en tercera persona del presente singular
WDT	Wh-determinador
WP	Wh-pronombre
WP$	Posesivo wh-pronombre
WRB	Wh-adverbio

Tabla A.3: Etiquetas sintácticas de SemCor

Apéndice B

WordNet

B.1 Estructura de la base de datos de WordNet

La estructura de la base datos viene determinada por la división del léxico en cuatro categorías y por las relaciones léxicas tratadas en la sección 3.3.1. Describimos a continuación la organización de la base de datos y de los principales ficheros que la constituyen. La estructura de la base de datos y el formato concreto de sus ficheros se ha desarrollado con vistas a facilitar el acceso directo desde diferentes aplicaciones. Así, por ejemplo, todos los ficheros almacenan la información codificada en formato ASCII.

En la tabla B.1 aparece el conjunto de ficheros que constituyen la base de datos en su versión 1.6 con sus tamaños correspondientes. Los más importantes son los ficheros `*.data` e `*.idx`. Se utiliza la división en categorías sintácticas (nombres, verbos, adjetivos y adverbios) para almacenar la información correspondiente en diferentes ficheros (con comienzos correspondientes `noun.`, `verb.`, `adj.`, `adv.`). En la figura 3.5 se representa gráficamente la información existente en cada uno de ellos, que detallamos a continuación. En la organización de la base de datos representa un papel central el *synset*. En la base de datos cada *synset* es representado por un número entero que lo identifica unívocamente. En los ficheros de índices (`*.idx`) se almacena la información relativa a cada término y los *synsets* a los que se encuentra asociado cada uno de ellos. Por ejemplo, el término "car" se encuentra asociado con los *synsets* `[02383458, 02384604, 02384960, 02385109, 02364995]` correspondientes a cada uno de sus cinco posibles significados distintos. En estos ficheros se almacena la información correspondiente a cada término en una línea y se encuentran ordenados alfabéticamente para facilitar la búsqueda. En cada línea se incluyen, básicamente, el término, y los identificadores de los *synsets* correspondientes. Como ejemplo, en el fichero `noun.idx`, se pueden encontrar las líneas que se muestran en la figura B.1:

En la sexta línea aparece el término "car" y sus *synsets* asociados. En los ficheros de datos (`*.dat`) se almacena la información correspondiente a las diferentes relaciones léxicas (hiperonimia, meronimia, etc.) existente en la base de datos, que se encuentran como asociaciones o enlaces entre los *synsets*. En los ficheros, se encuentra la información asociada a cada concepto

Fichero	Tamaño (KBytes)
noun.dat	10.733
verb.dat	1.850
adj.dat	2820
adv.dat	499
noun.idx	3.691
verb.idx	457
adj.idx	751
adv.idx	162
gloss.idx	5.576
sense.idx	5.982
noun.exc	109
verb.exc	82
adj.exc	20
adv.exc	1
sentidx.vrb	59
setns.vrb	5

Tabla B.1: Ficheros de la base de datos léxica WordNet (v. 1.6)

o *synset* en una línea, por orden creciente para facilitar las búsquedas. Para cada *synset*, se almacenan aquellos otros con los que se encuentra asociado por las diferentes relaciones.

Por ejemplo, en el fichero `noun.dat` se puede encontrar información correspondiente a nueve *synsets* en la base de datos, como se muestra en la figura B.2. Uno de ellos, el `02383458` (en el tercer fragmento de texto de la figura) es el correspondiente a uno de los cinco significados del término "car". En esta línea se incluyen los identificadores de todos aquellos *synsets* a los que se encuentra asociado por las diferentes relaciones léxicas, algunas de las cuales pueden verse representadas en la figura 3.5

```
...
capuchin n 2 2 @ #m 2 0 02383379 01986122
capulin n 2 3 @ #p %p 2 0 08942748 05793637
capulin_tree n 1 2 @ %p 1 0 08942748
caput n 2 4 @ ~ #p %p 2 0 10012369 04290247
capybara n 1 1 @ 1 0 01866573
car n 5 5 @ ~ #m #p %p 5 2 02383458 02384604 02385109
  02384960 02364995
car-ferry n 1 1 @ 1 0 02388365
car-mechanic n 1 1 @ 1 0 07092754
car_battery n 1 2 @ #p 1 0 02385783
...
```

Figura B.1: Muestra del fichero “noun.idx”

```
02383136 06 n 02 captopril 0 Capoten 0 001 @ 02195598 n 0000 |
 a drug (trade name Capoten) that blocks the formation of
 angiotensin in the kidneys resulting in vasodilation; used in
 the treatment of hypertension and congestive heart failure

02383379 06 n 01 capuchin 0 001 @ 02451635 n 0000 | a hooded cloak
 for women

02383458 06 n 05 car 0 auto 0 automobile 0 machine 1 motorcar 0
 052 @ 03018224 n 0000 %p 02159201 n 0000 %p 02169773 n 0000 ~
 02181645 n 0000 %p 02224961 n 0000 %p 02226945 n 0000 %p 02227084
 n 0000 ~ 02269665 n 0000 %p 02315972 n 0000 %p 02345631 n 0000 %p
 02352536 n 0000 ~ 02356871 n 0000 ~ 02361877 n 0000 %p 02387891 n
 0000 %p 02389506 n 0000 %p 02396780 n 0000 ~ 02478218 n 0000 ~
 02495126 n 0000 ~ 02510373 n 0000 ~ 02527747 n 0000 %p 02589341 n
 0000 %p 02670992 n 0000 %p 02688619 n 0000 %p 02700869 n 0000 %p
 02745031 n 0000 %p 02757950 n 0000 %p 02772072 n 0000 ~ 02798681
 n 0000 ~ 02803468 n 0000 ~ 02810347 n 0000 %p 02818207 n 0000 %p
 02827297 n 0000 ~ 02833216 n 0000 ~ 02836803 n 0000 ~ 02874141 n
 0000 ~ 02926688 n 0000 ~ 03074108 n 0000 ~ 03199039 n 0000 %p
 03217511 n 0000 %p 03235964 n 0000 ~ 03245084 n 0000 %p 03250547 n
 0000 %p 03262928 n 0000 %p 03296309 n 0000 ~ 03297658 n 0000 ~
 03388461 n 0000 ~ 03417119 n 0000 %p 03443129 n 0000 %p 03465436 n
 0000 %p 03496773 n 0000 ~ 03522520 n 0000 %p 03620585 n 0000 |
 4-wheeled motor vehicle; usually propelled by an internal combustion
 engine; "he needs a car to get to work"
...
02385783 06 n 02 car_battery 0 automobile_battery 0 002 @ 03421149
 n 0000 #p 02625930 n 0000 | a storage battery in a car; the heart of
 the car's electrical system
...
```

Figura B.2: Muestra del fichero "noun.dat"

Apéndice C

Detalles adicionales de los experimentos

En este apéndice se incluyen una relación de los experimentos de resolución de la ambigüedad léxica realizados a través del enfoque basado en SemCor y del basado en WordNet.

C.1 Experimentos WSD basados en SemCor

En la tabla C.1 se muestra la *precision* mediante *microaveraging* y *macroaveraging*, correspondiente a la totalidad de los términos (polisémicas y monosémicas) de los 50 primeros documentos que integran el Brown1 de SemCor. Se ha realizado una evaluación cruzada, evaluando cada documento (colección de prueba) sobre el resto de la colección de documentos (colección de entrenamiento).

	Precision			Precision	
Docs	microavg	macroavg	Docs	microavg	macroavg
br-a01	70,16	71,49	br-f43	64,55	66,87
br-a02	76,25	78,95	br-g01	68,88	72,24
br-a11	82,14	83,33	br-g11	67,38	70,54
br-a12	73,76	75,62	br-g15	62,56	65,66
br-a13	75,62	76,34	br-h01	65,22	67,30
br-a14	80,45	82,36	br-j01	75,61	77,59
br-a15	75,90	78,06	br-j02	73,24	76,34
br-b13	74,77	74,96	br-j03	73,23	76,25
br-b20	69,30	71,18	br-j04	75,84	76,55
br-c01	69,91	70,93	br-j05	71,05	71,62
br-c02	73,22	76,61	br-j06	74,83	76,69
br-c04	73,66	74,77	br-j07	75,15	76,54
br-d01	61,83	64,77	br-j08	74,36	75,81
br-d02	67,70	70,81	br-j09	77,42	78,81
br-d03	62,57	64,68	br-j10	73,71	76,26
br-d04	60,87	65,48	br-j11	75,00	76,00
br-e01	62,60	64,72	br-j12	75,44	77,52
br-e02	70,00	74,10	br-j13	65,52	67,74
br-e04	66,84	69,01	br-j14	75,19	76,59
br-e21	74,62	76,40	br-j15	78,75	79,30
br-e24	70,62	72,12	br-j16	79,38	79,79
br-e29	68,53	72,13	br-j17	70,13	71,12
br-f03	64,45	67,54	br-j18	55,56	55,68
br-f10	70,06	72,99	br-j19	68,80	70,26
br-f19	73,66	75,56	br-j20	58,62	60,00

Tabla C.1: Resultados correspondientes a la evaluación de los 50 primeros documentos del Brown1 de SemCor

C.2 Experimentos WSD basados en WordNet

A continuación, se muestran los resultados detallados de los experimentos correspondientes a la totalidad de los términos (polisémicas y monosémicas) de los 50 primeros documentos del Brown1 de SemCor, correspondientes a las principales relaciones de WordNet:

- *sinonimia.* Tabla C.2
- *hiponimia.* Tabla C.3
- *hiperonimia.* Tabla C.4
- *meronimia.* Tabla C.5
- *holonimia.* Tabla C.6
- *antonimia.* Tabla C.7

En las tablas se incluye la *precision* mediante *microaveraging* y *macroaveraging.*

Relación *syns* (sinonimia)					
Docs	Microavg	Macroavg	Docs	Microavg	Macroavg
br-a01	100,00	100,00	br-f43	82,81	90,23
br-a02	97,99	97,99	br-g01	92,14	94,16
br-a11	100,00	100,00	br-g11	90,40	94,12
br-a12	94,34	96,15	br-g15	92,97	95,97
br-a13	100,00	100,00	br-h01	92,52	96,12
br-a14	98,58	99,29	br-j01	93,90	96,84
br-a15	95,92	97,92	br-j02	94,25	96,47
br-b13	96,67	97,75	br-j03	97,30	98,63
br-b20	97,44	98,70	br-j04	97,94	97,94
br-c01	98,73	99,36	br-j05	94,69	95,95
br-c02	92,72	95,24	br-j06	90,29	93,94
br-c04	98,26	99,12	br-j07	90,18	93,52
br-d01	81,00	89,14	br-j08	95,71	97,10
br-d02	90,24	93,67	br-j09	94,12	96,39
br-d03	89,84	94,63	br-j10	96,80	97,58
br-d04	80,39	89,56	br-j11	93,98	96,30
br-e01	97,89	98,94	br-j12	94,59	96,53
br-e02	93,66	97,08	br-j13	92,45	94,23
br-e04	91,82	94,39	br-j14	90,76	94,74
br-e21	89,83	92,98	br-j15	97,14	98,08
br-e24	93,14	95,00	br-j16	95,00	97,44
br-e29	92,50	95,69	br-j17	92,24	93,42
br-f03	92,16	94,95	br-j18	95,83	95,83
br-f10	97,76	98,50	br-j19	90,91	95,24
br-f19	92,74	94,26	br-j20	85,71	92,31

Tabla C.2: Experimentos correspondientes a los primeros 50 documentos del Brown1 mediante la relación *sinonimia* de WordNet

Relación *hypo* (hiponimia)					
Docs	Microavg	Macroavg	Docs	Microavg	Macroavg
br-a01	98,95	98,95	br-f43	93,75	95,04
br-a02	97,73	97,73	br-g01	93,88	94,85
br-a11	98,61	98,61	br-g11	93,91	93,91
br-a12	98,21	98,21	br-g15	96,52	96,93
br-a13	98,33	98,33	br-h01	96,43	96,99
br-a14	100,00	100,00	br-j01	91,67	93,62
br-a15	96,55	96,55	br-j02	96,05	96,05
br-b13	96,92	96,92	br-j03	94,74	94,74
br-b20	97,85	97,85	br-j04	94,87	94,87
br-c01	100,00	100,00	br-j05	95,00	95,00
br-c02	97,22	98,43	br-j06	96,43	97,27
br-c04	98,82	98,82	br-j07	94,94	95,51
br-d01	92,31	92,86	br-j08	99,07	99,07
br-d02	95,59	95,59	br-j09	91,67	92,55
br-d03	96,55	96,55	br-j10	95,92	95,92
br-d04	92,65	94,44	br-j11	89,80	91,49
br-e01	100,00	100,00	br-j12	96,23	96,23
br-e02	98,88	98,88	br-j13	88,89	88,89
br-e04	94,12	94,12	br-j14	93,75	93,75
br-e21	97,85	98,37	br-j15	98,96	98,96
br-e24	94,81	95,39	br-j16	100,00	100,00
br-e29	95,96	97,25	br-j17	95,29	95,83
br-f03	94,25	94,25	br-j18	91,67	91,67
br-f10	97,00	97,00	br-j19	91,84	93,62
br-f19	95,65	95,65	br-j20	93,55	95,00

Tabla C.3: Experimentos correspondientes a los primeros 50 documentos del Brown1 mediante la relación *hiponimia* de WordNet

Relación *hype* (hiperonimia)					
Docs	Microavg	Macroavg	Docs	Microavg	Macroavg
br-a01	98,48	98,48	br-f43	75,58	78,02
br-a02	98,01	98,01	br-g01	89,03	90,40
br-a11	99,21	99,21	br-g11	75,58	77,54
br-a12	97,09	97,55	br-g15	89,05	89,05
br-a13	99,21	99,21	br-h01	87,60	88,19
br-a14	99,30	99,30	br-j01	82,47	84,57
br-a15	98,95	98,95	br-j02	83,33	84,03
br-b13	96,81	97,33	br-j03	76,79	77,98
br-b20	93,49	93,49	br-j04	83,59	84,25
br-c01	96,32	96,32	br-j05	88,19	88,19
br-c02	96,53	96,53	br-j06	88,98	90,09
br-c04	99,15	99,15	br-j07	79,73	80,48
br-d01	71,31	73,30	br-j08	91,19	91,19
br-d02	81,82	82,99	br-j09	92,00	92,42
br-d03	92,97	94,05	br-j10	90,34	90,34
br-d04	79,67	83,63	br-j11	88,17	88,17
br-e01	94,85	94,85	br-j12	92,77	93,29
br-e02	93,96	94,59	br-j13	81,82	81,82
br-e04	86,40	87,30	br-j14	89,92	90,62
br-e21	80,67	81,63	br-j15	95,83	96,28
br-e24	89,66	89,66	br-j16	92,05	92,53
br-e29	84,56	84,80	br-j17	79,01	79,58
br-f03	84,09	85,27	br-j18	81,82	84,38
br-f10	96,40	96,74	br-j19	92,31	92,19
br-f19	81,44	82,02	br-j20	89,13	90,00

Tabla C.4: Experimentos correspondientes a los primeros 50 documentos del Brown1 mediante la relación *hiperonimia* de WordNet

Relación *meron* (meronimia)					
Docs	Microavg	Macroavg	Docs	Microavg	Macroavg
br-a01	100,00	100,00	br-f43	96,97	96,97
br-a02	100,00	100,00	br-g01	96,97	96,97
br-a11	100,00	100,00	br-g11	94,12	95,92
br-a12	96,97	98,44	br-g15	100,00	100,00
br-a13	100,00	100,00	br-h01	100,00	100,00
br-a14	100,00	100,00	br-j01	100,00	100,00
br-a15	100,00	100,00	br-j02	100,00	100,00
br-b13	100,00	100,00	br-j03	100,00	100,00
br-b20	97,56	97,56	br-j04	100,00	100,00
br-c01	100,00	100,00	br-j05	100,00	100,00
br-c02	97,37	97,37	br-j06	100,00	100,00
br-c04	100,00	100,00	br-j07	100,00	100,00
br-d01	100,00	100,00	br-j08	100,00	100,00
br-d02	100,00	100,00	br-j09	100,00	100,00
br-d03	100,00	100,00	br-j10	100,00	100,00
br-d04	100,00	100,00	br-j11	100,00	100,00
br-e01	96,00	96,00	br-j12	100,00	100,00
br-e02	97,50	98,72	br-j13	92,86	92,86
br-e04	95,24	95,24	br-j14	100,00	100,00
br-e21	95,45	97,67	br-j15	100,00	100,00
br-e24	100,00	100,00	br-j16	100,00	100,00
br-e29	100,00	100,00	br-j17	100,00	100,00
br-f03	96,30	96,30	br-j18	100,00	100,00
br-f10	96,43	96,43	br-j19	100,00	100,00
br-f19	96,15	96,15	br-j20	100,00	100,00

Tabla C.5: Experimentos correspondientes a los primeros 50 documentos del Brown1 mediante la relación *meronimia* de WordNet

Relación *holon* (holonimia)					
Docs	Microavg	Macroavg	Docs	Microavg	Macroavg
br-a01	100,00	100,00	br-f43	100,00	100,00
br-a02	100,00	100,00	br-g01	100,00	100,00
br-a11	98,25	98,25	br-g11	100,00	100,00
br-a12	100,00	100,00	br-g15	95,24	95,24
br-a13	100,00	100,00	br-h01	100,00	100,00
br-a14	100,00	100,00	br-j01	00,00	100,00
br-a15	100,00	100,00	br-j02	100,00	100,00
br-b13	100,00	100,00	br-j03	100,00	100,00
br-b20	100,00	100,00	br-j04	100,00	100,00
br-c01	100,00	100,00	br-j05	100,00	100,00
br-c02	100,00	100,00	br-j06	100,00	100,00
br-c04	100,00	100,00	br-j07	0,00	100,00
br-d01	100,00	100,00	br-j08	100,00	100,00
br-d02	100,00	100,00	br-j09	100,00	100,00
br-d03	100,00	100,00	br-j10	98,04	98,04
br-d04	100,00	100,00	br-j11	100,00	100,00
br-e01	100,00	100,00	br-j12	100,00	100,00
br-e02	100,00	100,00	br-j13	100,00	100,00
br-e04	100,00	100,00	br-j14	100,00	100,00
br-e21	96,77	98,33	br-j15	100,00	100,00
br-e24	100,00	100,00	br-j16	100,00	100,00
br-e29	100,00	100,00	br-j17	100,00	100,00
br-f03	100,00	100,00	br-j18	100,00	100,00
br-f10	100,00	100,00	br-j19	100,00	100,00
br-f19	100,00	100,00	br-j20	100,00	100,00

Tabla C.6: Experimentos correspondientes a los primeros 50 documentos del Brown1 mediante la relación *holonimia* de WordNet

Relación *ants* (antonimia)					
Docs	Microavg	Macroavg	Docs	Microavg	Macroavg
br-a01	100,00	100,00	br-f43	95,45	97,62
br-a02	100,00	100,00	br-g01	100,00	100,00
br-a11	100,00	100,00	br-g11	100,00	100,00
br-a12	100,00	100,00	br-g15	100,00	100,00
br-a13	100,00	100,00	br-h01	100,00	100,00
br-a14	100,00	100,00	br-j01	100,00	100,00
br-a15	100,00	100,00	br-j02	100,00	100,00
br-b13	100,00	100,00	br-j03	100,00	100,00
br-b20	100,00	100,00	br-j04	100,00	100,00
br-c01	100,00	100,00	br-j05	100,00	100,00
br-c02	100,00	100,00	br-j06	100,00	100,00
br-c04	100,00	100,00	br-j07	100,00	100,00
br-d01	100,00	100,00	br-j08	100,00	100,00
br-d02	100,00	100,00	br-j09	100,00	100,00
br-d03	100,00	100,00	br-j10	100,00	100,00
br-d04	100,00	100,00	br-j11	100,00	100,00
br-e01	100,00	100,00	br-j12	100,00	100,00
br-e02	100,00	100,00	br-j13	100,00	100,00
br-e04	100,00	100,00	br-j14	100,00	100,00
br-e21	100,00	100,00	br-j15	100,00	100,00
br-e24	100,00	100,00	br-j16	100,00	100,00
br-e29	100,00	100,00	br-j17	100,00	100,00
br-f03	100,00	100,00	br-j18	100,00	100,00
br-f10	100,00	100,00	br-j19	100,00	100,00
br-f19	100,00	100,00	br-j20	100,00	100,000

Tabla C.7: Experimentos correspondientes a los primeros 50 documentos del Brown1 mediante la relación *antonimia* de WordNet

Apéndice D

Lista de Topics y expresiones textuales asociadas

En este apéndice se presentan el conjunto de *Topics* utilizados en la clasificación de la colección Reuters-21578, así como la relación de *Topics* empleados en la desambiguación, juntamente con las expresiones textuales con las que se les referencian en los cuerpos de los documentos de la colección.

N	Topic	N	Topic
1	ACQ	35	DRACHMA
2	ALUM	36	EARN
3	AUSTDLR	37	ESCUDO
4	AUSTRAL	38	F-CATTLE
5	BARLEY	39	FFR
6	BFR	40	FISHMEAL
7	BOP	41	FLAXSEED
8	CAN	42	FUEL
9	CARCASS	43	GAS
10	CASTOR-MEAL	44	GNP
11	CASTOR-OIL	45	GOLD
12	CASTORSEED	46	GRAIN
13	CITRUSPULP	47	GROUNDNUT
14	COCOA	48	GROUNDNUT-MEAL
15	COCONUT	49	GROUNDNUT-OIL
16	COCONUT-OIL	50	HEAT
17	COFFEE	51	HK
18	COPPER	52	HOG
19	COPRA-CAKE	53	HOUSING
20	CORN	54	INCOME
21	CORN-OIL	55	INSTAL-DEBT
22	CORNGLUTENFEED	56	INTEREST
23	COTTON	57	INVENTORIES
24	COTTON-MEAL	58	IPI
25	COTTON-OIL	59	IRON-STEEL
26	COTTONSEED	60	JET
27	CPI	61	JOBS
28	CPU	62	L-CATTLE
29	CRUDE	63	LEAD
30	CRUZADO	64	LEI
31	DFL	65	LIN-MEAL
32	DKR	66	LIN-OIL
33	DLR	67	LINSEED
34	DMK	68	LIT

Tabla D.1: Lista de *Topics* utilizados en la categorización (I)

N	Topic	N	Topic
69	LIVESTOCK	103	RINGGIT
70	LUMBER	104	RUBBER
71	LUPIN	105	RUPIAH
72	MEAL-FEED	106	RYE
73	MEXPESO	107	SAUDRIYAL
74	MONEY-FX	108	SFR
75	MONEY-SUPPLY	109	SHIP
76	NAPHTHA	110	SILK
77	NAT-GAS	111	SILVER
78	NICKEL	112	SINGDLR
79	NKR	113	SKR
80	NZDLR	114	SORGHUM
81	OAT	115	SOY-MEAL
82	OILSEED	116	SOY-OIL
83	ORANGE	117	SOYBEAN
84	PALLADIUM	118	STG
85	PALM-MEAL	119	STRATEGIC-METAL
86	PALM-OIL	120	SUGAR
87	PALMKERNEL	121	SUN-MEAL
88	PESETA	122	SUN-OIL
89	PET-CHEM	123	SUNSEED
90	PLATINUM	124	TAPIOCA
91	PLYWOOD	125	TEA
92	PORK-BELLY	126	TIN
93	POTATO	127	TRADE
94	PROPANE	128	TUNG
95	RAND	129	TUNG-OIL
96	RAPE-MEAL	130	VEG-OIL
97	RAPE-OIL	131	WHEAT
98	RAPESEED	132	WOOL
99	RED-BEAN	133	WPI
100	RESERVES	134	YEN
101	RETAIL	135	ZINC
102	RICE		

Tabla D.2: Lista de *Topics* utilizados en la categorización (y II)

TOPICS	**Expresión asociada**
ACQ	Mergers
AUSTDLR	Australian Dollar
BFR	Belgian Franc
BOP	Balance of payments
CAN	Canadian dollar
CARCASS	Carcass
CASTOR-OIL	Castor oil
COCOA	Cocoa
COCONUT-OIL	Coconut oil
COFFEE	Coffee
CORN	Corn
CORN-OIL	Corn oil
COTTON	Cotton
COTTONSEED	Cottonseed
CPI	Consumer Price Index
CRUDE	Crude Oil
DKR	Danish Krone
DRACHMA	Drachma
EARN	Earnings
ESCUDO	Escudo
FFR	French Franc
FLAXSEED	Flaxseed
FUEL	Fuel Oil
GAS	Gasoline
GNP	Gross National Product
GRAIN	Grain
GROUNDNUT	Groundnut
HEAT	Heating Oil
HK	Hong Kong Dollar
HOG	Hog
INCOME	Personal Income
INSTAL-DEBT	Consumer Credit
INTEREST	Interest rates
INVENTORIES	Inventories
JET	Kerosene
JOBS	Unemployment
LIN-OIL	Linseed oil
LINSEED	Linseed
LIT	Italian lira
LIVESTOCK	Livestock
LUMBER	Lumber
LUPIN	Lupin
MEXPESO	Mexican Peso

Tabla D.3: Lista de *Topics* utilizados en la desambiguación (I)

TOPICS	**Expresión asociada**
MONEY-FX	Foreign Exchange
MONEY-SUPPLY	Money supply
NAPHTHA	Naphtha
NAT-GAS	Natural Gas
NKR	Norwegian Krone
NZDLR	New Zealand Dollar
OILSEED	Oilseed
ORANGE	Orange
PALLADIUM	Palladium
PALM-OIL	Palm oil
PESETA	Peseta
PET-CHEM	Petro-Chemicals
PLATINUM	Platinum
PLYWOOD	Plywood
PORK-BELLY	Pork belly
POTATO	Potato
PROPANE	Propane
RAND	Rand
RAPE-OIL	Rape oil
RAPESEED	Rapeseed
RESERVES	Reserve
RETAIL	Retail
RINGGIT	Ringgit
RUBBER	Rubber
RUPIAH	Rupiah
SAUDRIYAL	Saudi Arabian Riyal
SFR	Swiss Franc
SHIP	Shipping
SILK	Silk
SINGDLR	Singapore Dollar
SKR	Swedish Krona
SOYBEAN	Soybean
STG	Sterling
SUGAR	Sugar
TAPIOCA	Tapioca
TRADE	Trade
TUNG	Tung
TUNG-OIL	Tung oil
WHEAT	Wheat
WOOL	Wool
WPI	Wholesale Price Index
ZINC	Zinc

Tabla D.4: Lista de *Topics* utilizados en la desambiguación (y II)

Bibliografía

E. Agirre, X. Arregi, X. Artola, A. Díaz, K. Sarasola, y A. Soroa. Un diccionario activo vasco-castellano en un entorno de escritura. In *VI Simposio Internacional de Comunicación Social*, 1995.

E. Agirre y G. Rigau. Word sense disambiguation using conceptual density. In *Proceedings of COLING*, 1996.

K. Aijmer y B. Altenberg. *Introduction.* K. Aijmer and B. Altenberg (eds.), 1991.

J. Allen. *Natural Language Understanding.* Benjamin/Cummings, 1994.

R. Amsler. Research toward the development of a lexical knowledge base for natural language processing. In *Proceedings of SIGIR'89, ACM Press*, 1989.

B.T. Atkins. Semantic ID tags: Corpus evidence for dictionary senses. In *Proceedings of the Third Annual Conference of the UW Center for the New OED*, 1987.

S. Atkins. Tools for computer-aided lexicography: the HECTOR Project. In *Papers in Computational Lexicography: COMPLEX*, 1993.

S. Atkins, J. Clear, y N. Ostler. Corpus design criteria. *Literary and Linguistic Computing*, (7), 1992.

R. Baeza-Yates y B. Ribeiro-Neto. *Modern Information Retrieval.* ACM Press Books, New York, 1999.

N. Belkin y B. Croft. Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM*, 35(12), 1992.

D. Biber. Representativeness in corpus design. *Computational Linguistics*, 19(2), 1993.

W. Black. An experiment in computational discrimination of english word senses. *IBM Journal of Research and Development*, 32(2), 1988.

B. Boguraev y T. Briscoe. *Computational Lexicography for Natural Language Processing.* Longman, 1989.

B. Boguraev y J. Pustejovsky. Lexical ambiguity and the role of knowledge representation in lexicon design. In *Proceedings of the 13th International Conference on Computational Linguistics, COLING'90*, 1990.

B. Boguraev y J. Pustejovsky. *Corpus Processing for Lexical Acquisition*. E. by B. Boguraev and J. Pustejovsky, The MIT Press, 1996.

P. Bollmann. The normalized *recall* and related measures. In *Proceedings of SIGIR'83, ACM Press*, 1983.

L. Braden-Harder. Sense disambiguation using on-line dictionaries. In *Natural Language Processing: The PLNLP Approach. Kluwer Academic Publishers*, 1993.

P. Brown, S. Della Pietra, V. Della Pietra, y R. Mercer. Word sense diambiguation using statistical methods. In *Proceedings of ACL*, 1991.

R. Bruce y W. Janyce. Word sense disambiguation using decomposable models. In *Proceedings of 33rd Annual Meeting of the Association for Computational Linguistics (ACL'94)*, 1994.

R. Bruce y J. Wiebe. A new approach to word sense disambiguation. In *Proceedings ARPA*, 1994a.

R. Bruce y J. Wiebe. Word sense disambiguation using decomposable models. In *Proceedings of ACL*, 1994b.

C. Buckley. Implementation of the SMART Information Retrieval System. Technical Report 85-686, Cornell University, 1985.

M. Buenaga. *Integración de Técnicas del Procesamiento del Lenguaje Natural para la Recuperación de Información en Bibliotecas de Componentes Software*. PhD thesis, Departamento de Informática y Automática. Universidad Complutense de Madrid, 1996.

M. Buenaga, J.M. Gómez, y B. Díaz. Using WORDNET to complement training information in text categorization. In *Proceedings of Second International Conference on Recent Advances in Natural Language Processing (RANLP)*, 1997.

L. Burnard. *The Text Encoding Initiative: a Progress Report*. G. Leitner (ed.), 1992.

L. Burnard. *User's Reference Guide for the British National Corpus. Version 1.0*. Oxford University Computing Services, 1995.

T. Carroll y C. Grover. *The Alvey Natural Language Tools Grammar (4th Release)*. Human Communication Research Centre, University of Edinburgh, 1993.

Y.K. Chang, C. Cirillo, y J. Razon. *Evaluation of Feedback Retrieval Using Modified Freezing, Residual Collection, and Test and Control Groups*. In, Salton [1971], 1971.

N. Chomsky. *Syntactic Structures.* Mouton, 1957.

K. W. Church y R. L. Mercer. *Introduction to the Special Issue on Computational Linguistics Using Large Corpora.* E. by Susan Armstrong In Using Large Corpora, 1993.

G.W. Cottrell. *A Connectionist Approach to Word Sense Disambiguation.* Pitman, 1989.

J. Cowie, J. Guthrie, y L. Guthrie. Lexical disambiguation using simulated annealing. In *Proceedings of the 14th International Conference on Computational Linguistics, COLING'92*, 1992.

J. Cowie y W. Lehnert. Information extraction. *Communications of the ACM*, 39(1), 1996.

B. Croft. The use of phrases and structured queries in information retrieval. In *Proceedings of SIGIR'91, ACM Press*, 1991.

B. Croft. Knowledge-based and statistical approaches to text retrieval. *IEEE Expert*, 8(2), 1993.

I. Dagan y A. Itai. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4), 1994.

K.L. Daghlgren. *Naive Semantics for Natural Language Understanding.* Kluwer Academic Publisher, 1988.

Daspa. *Discriminador Automático del Sentido de las Palabras en Textos Escritos.* Proyecto fin de carrera. Universidad de Jaén. Realizado por J.J. Contreras. Dirigido por L.A. Ureña, 1999.

M. Davies. The polyglot bible. In *http://mdavies.for.ilstu.edu/polyglot/*, 1999.

A. Díaz, M. Buenaga, L.A. Ureña, y M. García. Integrating linguistic resources in an uniform way for text classification tasks. In *Proceedings of the First International Conference on Language Resources and Evaluation —LREC98—*, 1998.

L.L. Earl. Use of word government in resolving syntactic and semantic ambiguities. *Information Storage and Retrieval*, (9), 1973.

J. Edwards y M. Lampert. *Talking Data: Transcription and Coding in Discourse Research.* Lawrence Erlbaum Associates Publishers, 1992.

S.P. Engelson y I. Dagan. Minimizing manual annotation cost in supervised training from corpora. In *http://xxx.lanl.gov/ps/cmp-lg/9606030*, 1996.

Cdad. Europea. *Lenguaje y tecnología. De la torre de Babel a la aldea global.* Oficina de Publicaciones Oficiales de las Comunidades Europeas, 1997.

C. Fellbaum. WordNet*: An Electronic Lexical Database.* E. by C. Fellbaum, The MIT Press, 1998.

J. Firth. A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis, Philological Society*, 1957.

C. Fox. *Lexical Analysis and Stoplists*, chapter 7. In, Frakes y Baeza [1992], 1992.

W. Frakes. *Stemming Algorithms*, chapter 8. In, Frakes y Baeza [1992], 1992.

W. Frakes y R. Baeza. *Information Retrieval: Data Structures and Algorithms.* Prentice-Hall, 1992.

W. Francis. *Problems of Assembling and Computerizing Large Corpora.* 1982.

W. Francis y H. Kucera. Frequency analisys of english usage. In *Houghton Mifflin*, 1982.

D.A. Gachot, E. Lange, y J. Yang. *The SYSTRAN NLP Browser: an application of Machine Translation Technology in Cross-Language Information Retrieval*, chapter 9. 1998.

W. Gale y K. Church. A program for aligning sentences in bilingual corpora. In *Proceedings of the ACL'91*, 1991.

W. Gale, K. Church, y D. Yarowsky. Work on statiscal methods for word snese disambiguation, in probabilistic approaches to natural language workshop. In *AAAI Fall Symposium*, 1992a.

W. Gale, K Church, y D. Yarowsky. A method for disambiguating word senses in a large corpus. In *Computers and the Humanities*, 1993.

W. Gale, K.W. Church, y D. Yarowsky. Estimating upper and lower bounds on the performance or word-sense disambiguation programs. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL'92)*, 1992b.

J. M. Gómez, A. Díaz, L.A. Ureña, y M. García. Utilización y evaluación de la desambiguación en tareas de clasificación de texto. *Procesamiento del Lenguaje Natural*, (25), 1999.

J. Gonzalo, F. Verdejo, I. Chugur, y J. Cigarrán. Indexing with WordNet synsets can improve text retrieval. In *http://xxx.lanl.gov/ps/cmp-lg/9808002*, 1998a.

J. Gonzalo, F. Verdejo, C. Peters, y N. Calzolari. Applying EuroWordNet to cross-language text tetrieval. *Computers and the Humanities*, (2/3), 1998b.

G. Grefenstette. Use of syntactic context to produce term association lists for text retrieval. In *Proceedings of SIGIR'92, ACM Press*, 1992.

J. Guthrie, L. Guthrie, Y. Wilks, y H. Aidinejad. Subject-dependent co-ocurrence and word sense disambiguation. In *Proceedings of the Annual Meeting*, 1991.

L. Guthrie. The role of lexicons in natural language processing. *Communications of the ACM*, 39(1), 1996.

D. Harman. *Relevance Feedback and other Query Modification Techniques.* 1992a.

D. Harman. Relevance feedback revisited. In *Proceedings of the 15th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, 1992b.

D. Harman. The first TExt Retrieval Conference (TREC-1). *Information Processing and Management*, 29(4), 1993.

D. Harman y E.M. Voorhees. Proceedings of the Sixth Text Retrieval Conference —(TREC-6)—. In *National Institute of Standards and Technology*, 1997.

P. Hayes. *Some Association-based Techniques for Lexical Disambiguation by Machine.* PhD thesis, Départment de Mathématiques, Ecole polytechnique Fédérale de Lausanne, 1977.

P. Hayes y S.P. Weinstein. CONSTRUE/TIS: a system for content-based indexing of a database of news stories. In *Second Annual Conference on Innovative Applications of Artificial Intelligence*, 1990.

M. Hearst. *Context and Structure in Automated Full-Text Information Access.* PhD thesis, Computer Science Division, University of California at Berkeley, 1994.

M. A. Hearst. Noun homograph disambiguation using local context in large text corpora. In *Proceedings of the 7th conference of the Centre for the New OED and Text Research: Using Corpora*, 1991.

G. Hirst. *Semantic Interpretation and the Resolution of Ambiguity.* Cambridge University Press, 1987.

L. Hjemslev. *Prolegomena to a Theory of Language Processing.* Indiana University, 1953.

J. Holland. Genetic algorithms. *Scientific American*, July 1992.

D. Hull. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of SIGIR'93*, 1993.

D. Hull y G. Grefenstette. Querying across languages: A dictionary-bases approach to multilingual information retrieval. In *the 19th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, 1996.

E. Ide. *New Experiments in Relevance Feedback.* In, Salton [1971], 1971.

N. Ide y J. Veronis. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1), 1998.

M. Inderjeet y M.T. Maybury. *Advances in Automatic Text Summarization.* The MIT Press, Cambridge, Massachusetss, 1999.

Instituto-Cervantes. Informe sobre recursos lingüísticos para el español. In *http://www.cervantes.es*, 1999.

S. Janssen. Tracing cohesive relations in corpora samples using dictionary data. In *New Directions in English Language Corpora, G. Leitner editor*, 1992.

T. Järvinen. Annotating 200 million words: The bank of english project. In *Proceedings of 15th International Conference on Computational Linguistics (COLING)*, 1994.

E.E. Kelly y P.J. Stone. Computer recognition of english word senses. *North-Holland*, 1975.

A. Kilgarriff. Corpus word usages and dictionary word senses: What is the match? an empirical study. In *Proceedings of the 7th Conference, UW Centre for the New OED and Text Research Using Corpora*, 1991.

A. Kilgarriff. *Polysemy.* PhD thesis, University of Sussex. School of Cognitive and Computer Sciences, 1992.

A. Kilgarriff. Dictionary word sense distintions: an enquiry into their nature. *Computers and the Humanities*, 26, 1993a.

A. Kilgarriff. I don't believe in word sense. *Computers and the Humanities*, 31(2), 1993b.

A. Kilgarriff. Gold standard for evaluating word sense disambiguation programs. *Computer Speech and Language, Special Issue on Evaluation*, 12(3), 1998.

R. Krovetz. Viewing morphology as an inference process. In *Proceedings of SIGIR'93, ACM Press*, 1993.

R. Krovetz y W.B. Croft. Word sense disambigation using machine-readable dictionaries. In *Proceedings of ACM SIGIR*, 1989.

F.W. Lancaster. MEDLARS: Report on the evaluation of its operating efficiency. *American Documentation*, 20, 1969.

S. Landes, C. Leacock, y R. Tengi. *Building semantic concordances*, chapter 8. In, Fellbaum [1998], 1998.

C. Leacock, G. Towell, y E. Voorhees. *Towards building contextual representations of word senses using statistical models*, chapter 6. In, Boguraev y Pustejovsky [1996], 1996.

G. Leech. *The State of the Art in Corpus Linguistics.* E. by K. Aijmer and B. Altenberg, 1991.

L. Leech, R. Garside, y M. Bryand. *The Large-Scale Grammatical Tagging of Text: Experience with the British National Corpus.* E. by N. Oostdijk and P. Haan, 1994.

S. Lehmann, S. Oepen, S. Regnier-Prost, K.Ñetter, V. Lux, J. Klein, K. Falkedal, F. Fouvry, D. Estival, E. Dauphin, H. Campagnion, J. Baur, L. Balkan, y D. Arnold. Test Suites for Natural Language Processing —TSNLP—. In *Proceedings of 16th International Conference on Computational Linguistics*, 1996.

W. Lehnert y B. Sundheim. An evaluation of text analysis technologies. *AI magazine*, 12(3), 1991.

D. Lenat. A large-scale investment in knowledge infrastructure. *Communications of de ACM*, 38(11), 1995.

M. Lesk. Automatic sense disambiguation: How to tell a pine cone from an ice cream cone. In *Proceedings of the ACM SIGDOC Conference*, 1986.

D.D. Lewis. *Representation and Learning in Information Retrieval.* PhD thesis, Department of Computer and Information Science, University of Massachusetts, 1992.

D.D. Lewis. Reuters-21578 text categorization test collection, 1997. Distribution 1.0, —AT&T Labs-Research—.

D.D. Lewis, R.E. Schapire, J.P. Callan, y R. Papka. Training algorithms for linear text classifiers. In *Proceedings of the ACM SIGIR*, 1996.

E.D. Liddy y W. Paik. Statistically-guided word sense disambiguation. In *Proceedings of the AAAI Fall Symposium Series*, 1993.

H.P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2), 1958.

M. Maña, L.A. Ureña, y M. Buenaga. Tareas de análisis del contenido textual para la recuperación de información con realimentación. *Procesamiento del Lenguaje Natural*, (24), 2000.

Y. Maarek y D. Berry. An information retrieval approach for automatically constructing software libraries. *IEEE Trans. Software Engineering*, 17(8), 1991.

C.D. Manning y H. Schütze. *Foundations of Statistical Natural Language Processing.* The MIT Press. Cambridge, Massachusetts, 1999.

M.P. Marcus, M.A. Marcinkiewicz, y B. Santorini. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2), 1993.

A. Meillet. Linguistique historique et linguistique générale. 1(2), 1926.

G. Miller. WordNet: An on-line lexical database. In *An International Journal of Lexicography*, 1990.

G. Miller. WordNet: A lexical database for english. *Communications of the ACM*, 38(11), 1995.

G. Miller y G. Charles. Contextual correlates of semantic similarity. In *Language and Cognitive Processes*, 1991.

G. Miller, C. Leacock, T. Randee, y R. Bunker. A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, 1993.

R.J. Mooney. Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1996.

H.T. Ng y H.B. Lee. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL'96)*, 1996.

OEIL. *Ingeniería lingüística. Cómo aprovechar la fuerza del lenguaje.* DG XIII/E. Publicación electrónica en http://www2.echo.lu/langeng/es/broch/harness.html, 1998.

Corp. Oracle. Managing text with oracle8(tm) context cartridge. In *An Oracle Technical White Paper*, 1997.

L. Padró. *A Hybrid Environment for Syntax-Semantic Tagging.* PhD thesis, Departamento de Lenguajes y Sistemas Informáticos, Universidad Politécnica de Cataluña, 1997.

R. Passonneau y D. Litman. Intention-based segmentation: Human reliability and correlation with linguistic cues. In *Proceedings of ACL-93*, 1993.

P. Pedersen, R Bruce, y J. Wiebe. Sequential model selection for word sense disambiguation. In *Proceedings 5th ANLP*, 1997.

M.F. Porter. An algorithm for suffix stripping. *Program-automated library and information systems*, 3(14), 1980.

R. Pressman. *Ingeniería del Software. Un efoque práctico.* McGraw-Hill, 1997.

Y. Qiu y H. Frei. Concept based query expansion. In *Proceedings of SIGIR'93*, 1993.

V. Raghavan, P. Bollmann, y G. Jung. Retrieval system evaluation using *recall* and *precision*: Problems and answers. In *Proceedings of SIGIR'89*, 1989.

E. Rasmussen. *Clustering Algorithms*, chapter 16. In, Frakes [1992], 1992.

P. Resnik. Disambiguating Noun Groupings with respect to WordNet Senses. In *Proceedings of the Third Workshop on Very Large Corpora*, 1995a.

P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligent, IJCAI'95*, 1995b.

P. Resnik y D. Yarowsky. A perspective on word sense disambiguation methods and their evaluation. In M. Light, editor, *Tagging Text with Lexical Semantics: Why, What and How?, ACL SIGLEX*, 1997.

R. Richardson y A. Smeaton. Using WordNet in a knowledge-based approach to information retrieval. In *Proceedings of the BCS-IRSG Colloquium*, 1995.

G. Rigau, E. Agirre, y J. Atserias. Combining unsupervised lexical knowledge methods for word sense disambiguation. In *Proceedings of joint ACL/EACL 1997*, 1997.

C.J.van Rijsbergen. *Information Retrieval.* Butterworths, 1979.

C.J.van Rijsbergen. A new theorical framework for information retrieval. In *Proceedings of the 9th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, 1986.

J.J. Rocchio. *Relevance Feedback in Information Retrieval.* In, Salton [1971], 1971.

J. Rumbaugh. *Object-oriented modeling and design.* Prentice-Hall, 1991.

M. Sahami, editor. *Proceedings of the AAAI'98/ICML'98 Workshop on Learning for Text Categorization*, 1998.

G. Salton. *The* Smart *Retrieval System: Experiments in automatic document processing.* E. by G. Salton, Prentice-Hall, Inc., 1971.

G. Salton. *Automatic Text Processing: the transformation, analysis and retrieval of information by computer.* Addison Wesley, 1989.

G. Salton. The SMART document retrieval project. In *Proceedings of SIGIR'91, ACM Press*, 1991a.

G. Salton. The state of retrieval system evaluation. Technical Report 91-1206, Department of Computer Science, Cornell University, 1991b.

G. Salton y J. Allan. Approaches to passage retrieval in full text information systems. In *Proceedings of SIGIR'93, ACM Press*, 1993.

G. Salton y C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4), 1990.

G. Salton y M.J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

G. Salton, A. Wong, y C. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 1975.

G. Sampson. *SUSANNE a Deeply Analysed Corpus of American English*. 1992.

M. Sanderson. *Word sense disambiguation and information retrieval*. PhD thesis, Department of Computing Science, University of Glasgow, 1996.

J. Scholtes. *Neural Networks in Natural Language Processing and Information Retrieval*. PhD thesis, Universiteit van Amsterdam, 1993.

H. Schütze. Word sense disambiguation with sublexical representations. In *Proceedings of the 1992 AAAI Workshop on Statistically-based Natural Language Programming Techniques*, 1992.

H. Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24(1), 1998.

S. Scott y S. Matwin. Text classification using WordNet hypernyms. In *Proceedings of the COLING-ACL'98 Workshop in Usage of* WordNet *in Natural Language Processing Systems*, 1998.

M. Sheldon. Discover: A resource discovery system based on content routing. In *Third International World Wide Web Conference*, 1995.

H. Shütze y J. Pedersen. Information retrieval based on word senses. In *Proceedings of 4th Symposium on document Analysis and Information Retrieval*, 1995.

E. Siegel. Learning methods for combining linguistics indicators to classify verbs. In *Proceedings of 2nd Conference on Empirical Methods for Natural Language Processing, EMNLP 97*, 1997.

J. Sinclair. *Looking Up: An Account of the COBUILD Project in Lexical Computing*. Collins, 1987.

J. Sinclair. *Corpus, Concordance, Collocation*. Oxford University Press, 1991.

B. Slator. Sense and preference. *Computer and Mathematics with Applications*, 23, 1992.

B. M. Slator y Y. Wilks. Towards semantic structures from dictionary entries. In *Proceedings of the 2nd annual rocky mountain conference on Artificial Intelligence*, 1987.

S.L. Small y C. Rieger. Parsing and comprehending with word experts (a theory and its realization). In *Strategies for natural language, W. Lenhert and M. Ringle, editors LEA*, 1982.

A. Smeaton, F. Kelledy, y R. O'Donell. TREC-4 Experiments at Dublin City University: thresholding posting lists, query expansions with WordNet and POS tagging of spanish. In *Proceedings of TREC-4*, 1995.

A. Smeaton y A. Quigley. Experiments on using semantic distances between words in image captio retrieval. In *Proceedings of the 19th International Conference on Research and Development in IR*, 1996.

C. Souter. *Towards a Standard format for Parsed Corpora.* E. by J. Aarts, P. de Haan and N. Oostdijk, 1993.

K. Sparck-Jones. Reflections on TREC. *Information Processing and Management*, 31(3), 1995.

P. Srinivasan. *Thesaurus Construction*, chapter 9. In, Frakes [1992], 1992.

M. Sussna. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the Second International Conference on Information and Knowledge Management CIKM'93*, 1993.

F. Thérèse. A proposal for a task-based evaluation of text summarization systems. In *Proceedings of ACL/EACL Workshop on Intelligent Scalable Text Summarization*, 1997.

A. Tversky. Features of similarity. *Psychological Review*, 84(4), 1977.

L.A. Ureña, M. Buenaga, M. García, y J.M. Gómez. Integrating and evaluating WSD in the adaptation of a lexical database in text categorization task. In *Proceedings of the First Workshop on Text, Speech, Dialogue —TSD'98—*, 1998a.

L.A. Ureña, M. Buenaga, y J.M. Gómez. Using and Evaluating WSD in Information Retrieval. In *Conference of the International Quantitative Linguistics Association* —QUALICO 2000—, 2000a.

L.A. Ureña, M. Buenaga, y J.M. Gómez. Integrating linguistic resources in TC through WSD. *Computers and the Humanities*, 35(2), 2001.

L.A. Ureña, M. García, M. Buenaga, y J.M. Gómez. Resolución de la ambigüedad léxica mediante información contextual y el modelo del espacio vectorial. In *Séptima Conferencia de la Asociación Española para la Inteligencia Artificial –CAEPIA–*, 1997.

L.A. Ureña, M. García, M. Buenaga, y J.M. Gómez. Resolución automática de la ambigüedad léxica fundamentada en el modelo del espacio vectorial usando ventana contextual variable. In *Asociación Española de Lingüística Aplicada*, 1998b.

L.A. Ureña, M. García, J.M. Gómez, y A. Díaz. Integrando una base de datos léxica y una colección de entrenamiento para la desambiguación del sentido de las palabras. *Procesamiento del Lenguaje Natural*, (23), 1998c.

L.A. Ureña, J.M. Gómez, y M. Buenaga. Information retrieval by means of word sense disambiguation. In *Lecture Notes in Artificial Intelligence. Third International Workshop on Text, Speech and Dialogue —TSD'2000—*, volume 1902. Springer-Verlag, 2000b.

A. Vaquero y M. Buenaga. Aplicaciones de las bases de datos léxicas en la clasificación automática de textos. In *Curso de la XVIII Escuela de Verano de Informática*, 1996.

F. Verdejo. Comprensión del lenguaje natural: avances, aplicaciones y tendencias en procesamiento del lenguaje natural: fundamentos y aplicaciones. In *Documentación del curso de verano de 1994 de la Universidad Nacional de Educación a Distancia, Ávila*, 1994.

J. Veronis y N. Ide. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *Proceedings of the 13th International Conference on Computational Linguistics, COLING'90*, volume 2, 1990.

J. Veronis y N. Ide. An assessment of information automatically extracted from machine readable dictionaries. In *Proceedings of the Fifth Conference European Chapter of the Association for Computational Linguistics*, 1991.

J. Veronis y N. Ide. *Large Neural Networks for the Resolution of Lexical Ambiguity.* 1995.

E. Voorhees. Using WORDNET to disambiguate word senses for text retrieval. In *Proceedings of SIGIR'93*, 1993.

E.M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the ACM SIGIR*, 1994.

P. Vossen y otros. The Restructured Core WordNets in EuroWordNet: Subset1. Deliverable D014, D015, WP3, WP4. EuroWordNet LE2-4003, 1998.

D. Walker y R. Amsler. The use of machine-readable dictionaries in sublanguage analysis. In *Ralph Grishman and Richard Kittredge (eds.), Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, 1986.

D. Walker, A. Zampolli, y N. Calzolari. *Automating the Lexicon: Research and Practice in a Multilingual Environment.* Oxford University Press, 1995.

S. Weiss. Learning to disambiguate. *Information Storage and Retrieval*, 9, 1973.

B. Widrow y S. Sterns. *Adaptative Signal Processing.* Prentice-Hall, Englewood Cliffs, 1985.

E.D. Wiener, J. Pedersen, y A.S. Weigend. A neural network approach to topic spotting. In *Proceedings of the SDAIR*, 1995.

R. Wilensky. The UC Berkeley Digital Library Project: re-thinking scholarly Information Dissemmination and Use. In *Lecture Notes in Computer Science. European Conference on Research and Advanced Technology for Digital Libraries (ECDL'99)*, volume 1696. Springer-Verlag, 1999.

Y. Wilks. An artificial intelligence approach to machine translation. In *Roger Schank and Kenneth Colby Editors, Computers Models of Thought and Language. W. H. Freeman*, 1973.

Y. Wilks. Is Word Sense Disambiguation just one more NLP task? In *Proceedings of the SENSEVAL Conference*, 1998.

Y. Wilks, D. Fass, C. Guo, J. McDonald, T. Plate, y B. Slator. Providing machine tractable dictionary tools. In *Machine Translation*, 1990.

Y. Wilks y M. Stevenson. Combining independent knowledge sources for word sense disambiguation. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP)*, 1997.

Y. Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval Journal*, 1(1/2), 1999.

Y. Yang y X. Liu. A re-examination of text categorization methods. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1999.

D. Yarowsky. Word-sense disambiguation using statistical models of ROGET'S categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, 1992.

D. Yarowsky. One sense per collocation. In *Proceedings of ARPA Human Language Technology Workshop*, 1993.

D. Yarowsky. A comparison of corpus-based thecniques for restoring accents in spanish and french text. In *Proceedings of the 2nd Annual Workshop on Very Large Text Corpora*, 1994a.

D. Yarowsky. Decision list for lexical ambiguity resolution: Application to accent restoration in spanish and french. In *Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics (ACL'94)*, 1994b.

D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33th Annual Meeting of the Association for Computational Linguistics (ACL'95)*, 1995.

T. Yokoi. The EDR electronic dictionary. *Communications of the ACM*, 38(11), 1995.

E. Yourdon y L. Constantine. *Structured Design.* Yourdon Press (Prentice Hall), 1979.

A. Zampolli, N. Calzolari, y M.Palmer. *Current Issues in Computational Linguistics: In Honour of Don Walker.* Giardini Editori, Kluwer, 1994.

G.K. Zipf. *Human Behavior and the Principle of Least Effort.* Addison-Wesley, 1949.