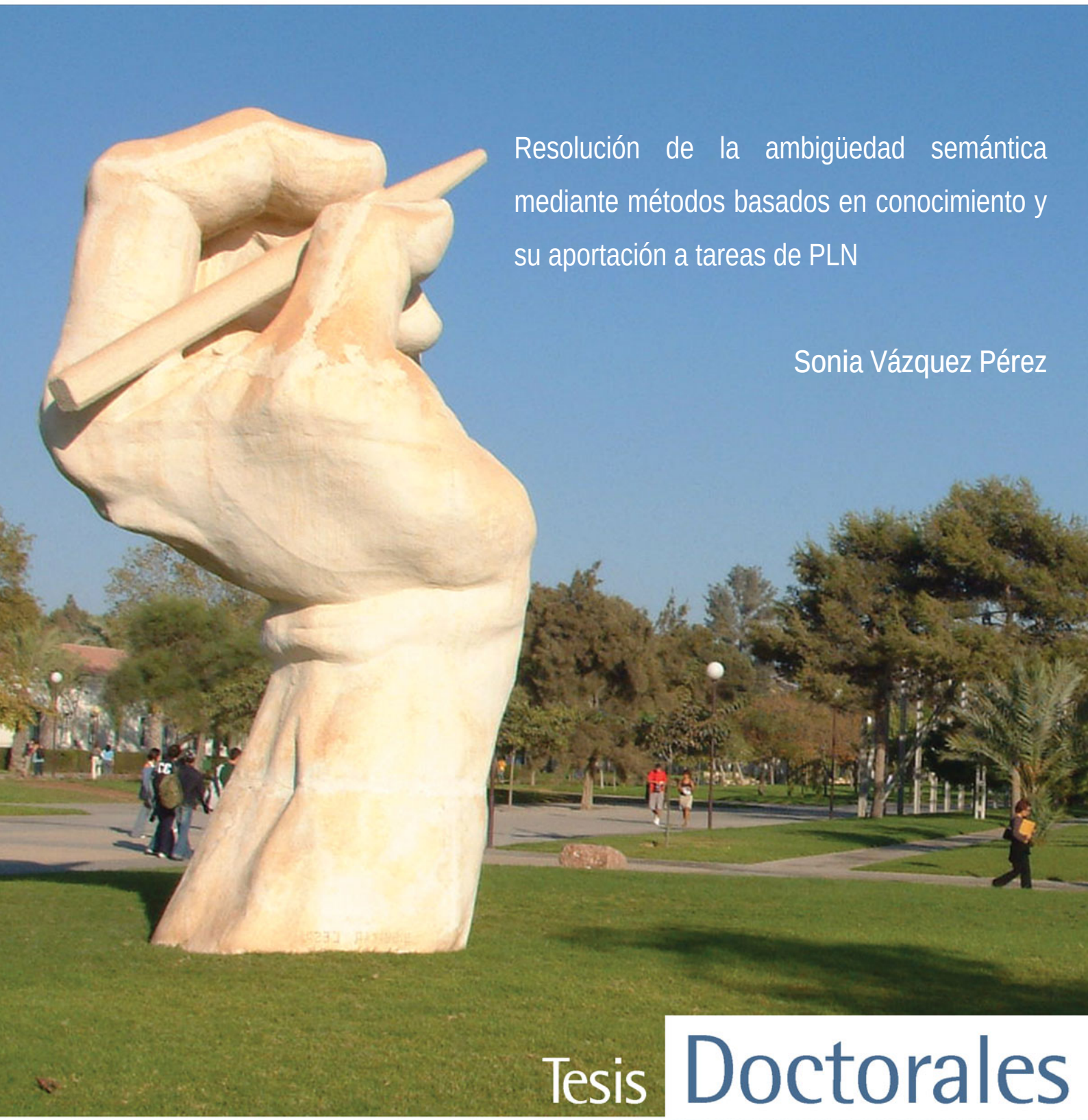
Universitat d'Alacant
Universidad de Alicante



Resolución de la ambigüedad semántica mediante métodos basados en conocimiento y su aportación a tareas de PLN

Sonia Vázquez Pérez

Tesis Doctorales

www.eltallerdigital.com

UNIVERSIDAD de ALICANTE

Universitat d'Alacant
Universidad de Alicante

Resolución de la ambigüedad semántica mediante métodos basados en conocimiento y su aportación a tareas de PLN

Tesis Doctoral

Autora: **Sonia Vázquez Pérez**

Director:
Dr. Andrés Montoyo Guijarro

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante

Alicante, 2009

Universitat d'Alacant
Universidad de Alicante

A mi familia

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Agradecimientos

En primer lugar me gustaría dar mi más sincero agradecimiento a todas las personas que me han animado y apoyado en la realización de esta Tesis. En especial, me gustaría agradecer a mi director, Andrés Montoyo, todo el tiempo que ha dedicado a dirigir mis investigaciones y todos los consejos que me ha dado durante toda mi trayectoria como investigadora, los cuales, me han llevado hasta este punto.

También quisiera hacer mención especial a todos mis compañeros del Grupo de Procesamiento del Lenguaje Natural de la Universidad de Alicante, que han aportado su granito de arena en esta Tesis y cuyos consejos han sido fundamentales para la finalización de este trabajo. Muchos de los avances de esta tesis han sido gracias a los trabajos realizados en colaboración con distintos miembros del grupo de donde han surgido muy buenas ideas.

Para terminar, no podría haber llegado hasta aquí sin el apoyo que toda mi familia me ha brindado durante todos estos años y sobretodo, gracias a mi marido por su comprensión y paciencia en este largo camino que sin él habría sido, sin duda, muy duro.

Alicante, 2009

Sonia Vázquez

Universitat d'Alacant
Universidad de Alicante

Índice general

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Índice de tablas

Universitat d'Alacant
Universidad de Alicante

Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Índice de figuras

Universidad de Alicante

Capítulo 1

Introducción

Desde la aparición de las primeras computadoras en los años 50, nuestras vidas giran en torno a multitud de dispositivos electrónicos que facilitan nuestra existencia. De hecho, en la actualidad, la gran explosión de tecnologías relacionadas con las comunicaciones (internet, telefonía móvil, etc) ha motivado el desarrollo parejo de otras tecnologías estrechamente vinculadas a mejorar la comunicación hombre-máquina. Actualmente, los dispositivos GPS utilizan sistemas de comunicación que simulan la voz humana, las búsquedas en internet se realizan en cualquier idioma, siendo el buscador capaz de reconocer el idioma y extraer la información correspondiente, las traducciones de textos se hacen de forma automática con un software especialmente diseñado para ello, etc. Un ejemplo de cómo han evolucionado los sistemas de comunicación son los actuales servicios de información automatizados, donde a través del teléfono podemos comprobar en breves segundos si existen atascos, encontrar la farmacia de guardia de un municipio, consultar el pronóstico meteorológico, etc. Todas las consultas se hacen como si realmente existiera una persona al otro lado, pero verdaderamente detrás del auricular hay un sistema automatizado muy complejo que procesa la consulta de forma automática. A simple vista todo parece muy sencillo, se hace la pregunta, se procesa y se da la respuesta. Pero de hecho, la parte de procesamiento lleva asociada una gran complejidad.

Concretamente, en el caso de las consultas telefónicas, para que todo funcione correctamente y se obtenga una respuesta satisfactoria, es necesario realizar previamente una serie de tareas más complejas como: determinar el idioma del interlocutor, transformar los sonidos en palabras y frases con significado, realizar el análisis sintáctico de la pregunta, detectar los nombres propios, seleccionar del sentido correcto de cada palabra dentro de su contexto, etc. Todas estas tareas requieren de un profundo conocimiento lingüístico y a la vez en muchos casos, de un elevado coste computacional. Estas necesidades han derivado en una disciplina denominada Lingüística Computacional o Procesamiento del Lenguaje Natural (PLN) que combina la lingüística y la informática con el fin de modelar el lenguaje humano desde un punto de vista computacional.

Uno de los motores principales que ha impulsado la necesidad actual del tratamiento del lenguaje humano ha sido Internet. La red proporciona una gran cantidad de información sobre multitud de temáticas pero con un problema asociado: la información está desestructurada y descentralizada. Existen infinidad de páginas web: de empresas, personales, blogs, foros, páginas web institucionales... algunas de ellas están relacionadas y otras no tienen nada en común. Esta falta de organización requiere la utilización de algún tipo de tecnología que gestione de forma eficaz toda la información disponible para que tanto las búsquedas como las consultas sean efectivas. Esta problemática ha derivado en la utilización de dos tipos de tecnologías bien diferenciadas: Tecnologías de Procesamiento de Datos y Tecnologías de Procesamiento del Lenguaje Natural. Cada una de estas tecnologías procesa de forma diferente la información. A diferencia de las Tecnologías de Procesamiento de Datos que se ocupan de reducir el espacio ocupado, almacenar de forma óptima los datos, ahorrar tiempos de respuesta en la búsqueda de algún tipo de información, etc, las Tecnologías de Procesamiento del Lenguaje Natural necesitan un conocimiento más profundo del lenguaje para poder procesar la información. Esta diferencia puede verse más clara con el siguiente ejemplo:

Supongamos que tenemos un programa que puede contar el número de líneas, de palabras o de bytes en un fichero de texto. Dentro de este programa se utilizan dos técnicas distintas. Mientras que para contar líneas y número de bytes no es necesario ningún conocimiento lingüístico (procesamiento de datos), para contar las palabras sí es necesario un conocimiento de qué significa realmente ser una palabra (procesamiento del lenguaje natural).

Este es un pequeño ejemplo muy simple y muy pobre en conocimiento lingüístico de aplicación de procesamiento del lenguaje natural. Por supuesto, actualmente se han desarrollado multitud de aplicaciones mucho más complejas que abordan diferentes tipos de problemas. Entre este tipo de aplicaciones encontramos: traductores automáticos, motores de búsqueda, sistemas de diálogo, etc. Todos estos sistemas deben profundizar mucho más en el conocimiento lingüístico para su correcto funcionamiento. Por ejemplo, un sistema de traducción automática debería ser capaz de traducir correctamente "*I'm in bed because I have a cold*" por "Estoy en cama porque estoy resfriado" y no por "Estoy en cama porque tengo un frío". La forma de determinar el significado de la palabra "*cold*" es utilizar las palabras del contexto que la rodean, para poder decidir qué sentido es el más apropiado. Por tanto, cuando necesitamos un conocimiento más preciso del lenguaje, de las relaciones entre palabras o de las expresiones en diferentes contextos, hablamos de Tecnologías de Procesamiento del Lenguaje Natural (PLN).

Universidad de Alicante

1.1 Problemática en PLN

Uno de los principales problemas encontrados al tratar texto no pre-formateado: diálogos, consultas telefónicas... es la ambigüedad. En el lenguaje humano podemos encontrar múltiples expresiones y palabras que pueden tener varios significados distintos dependiendo de las circunstancias de uso. Estas características hacen que el lenguaje natural se distinga de los lenguajes artificiales por su riqueza (en vocabulario y construcciones), flexibilidad (reglas con múltiples excepciones), ambigüedad (pudiendo darse

diversos significados de una palabra o una frase según el contexto), indeterminación (permitiendo referencias y elipsis) y distintas interpretaciones del sentido literal según la situación o el contexto en que se produce. Lo que son ventajas para la comunicación humana se convierten en problemas a la hora de un tratamiento computacional, ya que implican conocimiento y procesos de razonamiento que son difíciles de formalizar. Debido a estas características del lenguaje natural, es necesario utilizar una serie de técnicas de PLN para trabajar con texto y expresiones humanas que permitan resolver a partir de un análisis dirigido por el dominio, el uso, el contexto, etc, los distintos problemas de interpretación que puedan aparecer.

En el campo del PLN el problema de la ambigüedad puede tratarse desde distintas perspectivas. Desde la ambigüedad debida a palabras polisémicas, hasta la ambigüedad producida por las distintas interpretaciones que pueda tener una oración. Dentro del PLN, por tanto, podemos distinguir tres tipos de ambigüedad:

Ambigüedad léxica Una misma palabra puede pertenecer a diferentes categorías gramaticales.
Por ejemplo:
La palabra "para" puede ser: preposición, forma del verbo parar o forma del verbo parir.

Ambigüedad sintáctica o ambigüedad estructural Aparece cuando debido a la forma en que se asocian los distintos constituyentes de una oración, podemos interpretarla de varias formas distintas. Siendo a veces casi imposible de solucionar.
Por ejemplo:
Juan vio a su hermana con unos prismáticos (¿Juan usó los prismáticos para ver a su hermana o Juan vio que su hermana tenía unos prismáticos?)

Ambigüedad semántica Dentro de este tipo de ambigüedad podemos diferenciar tres clases:

1. Ambigüedad debida a las palabras polisémicas. En este caso, una misma palabra puede tener distintos significados dependiendo del uso que se le esté dando en cada momento.

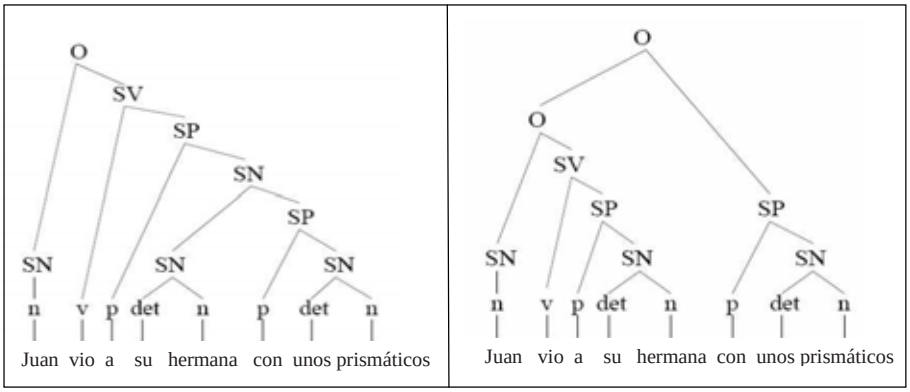


Figura 1.1. Ambigüedad sintáctica

Por ejemplo:
Entró en el **banco**. (Entidad financiera)
Se sentó en el **banco** del parque. (Asiento)

2. Ambigüedad debida a encontrar una misma estructura sintáctica con diferentes significados.
   Por ejemplo:
   Todos los estudiantes de secundaria **hablan dos lenguas**. (¿Cada estudiante habla dos lenguas o sólo se hablan dos lenguas determinadas?)
3. Ambigüedad referencial. En este caso, es necesario el análisis del texto más allá de los límites de la frase, determinando los antecedentes referenciales de los pronombres.
   Por ejemplo:
   El jamón está en el armario. **Sácalo**. **Ciérralo**.(¿Hacen referencia al jamón o al armario?)

La resolución de los diferentes tipos de ambigüedades requiere mucho conocimiento y es necesario aplicar diferentes técnicas para solucionar cada caso. Podemos utilizar modelos de Markov (Markov (1971)) para resolver la ambigüedad léxica, gramáticas probabilísticas para resolver la ambigüedad sintáctica o técnicas basadas en conocimiento o aprendizaje automático para resolver la ambigüedad semántica. El tratamiento de la ambigüedad, es por tanto, una tarea necesaria para cualquier sistema de PLN,

pero esta tarea no funciona de forma independiente, se complementa con otras tareas como el análisis sintático que le suministra información muy valiosa. De esta forma, podemos decir que la tarea de resolución de la ambigüedad es una tarea intermedia que completa un sistema de PLN.

1.2 Estructura de un sistema de PLN

Si se analizan en profundidad los actuales sistemas de PLN todos ellos comparten una serie de módulos básicos para su correcto funcionamiento. La Figura 1.2 muestra la estructura general de un sistema de PLN.

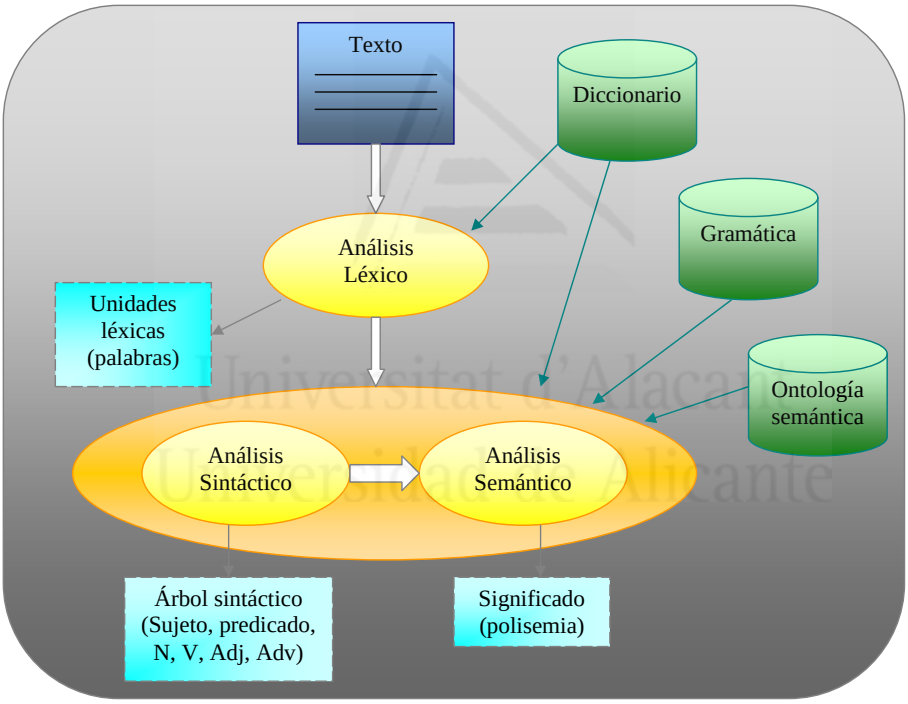


Figura 1.2. Estructura de un sistema de PLN

Como se puede apreciar en la Figura 1.2 y suponiendo que estamos procesando texto, tenemos tres módulos principales:

Módulo de análisis léxico. La principal tarea de este módulo es detectar palabras, es decir, la menor unidad existente con significado. Dentro del léxico de un lenguaje es necesario detectar además de las palabras simples, las palabras compuestas, frases hechas, siglas, préstamos idiomáticos, etc. Además también es necesario diferenciar entre la forma (la palabra tal como aparece) y el lema (la forma canónica de la palabra). El objetivo final de este módulo es asociar a cada palabra su lema correspondiente, etiquetar cada palabra con sus posibles categorías léxicas (N, V, Adj, Adv) y añadir algunos rasgos gramaticales (género, número, tiempo verbal...)

Módulo de análisis sintáctico. Este módulo se ocupa de realizar el análisis sintáctico de tal forma que selecciona la etiqueta gramatical más apropiada para cada palabra, realiza un "chunking" del texto (divide el texto en segmentos analizables), utiliza formas lógicas para el análisis, etc.

Módulo de análisis semántico. El último módulo es el que sigue al análisis sintáctico. En este caso se ocupa de asignar el sentido correspondiente a cada palabra (resolver la ambigüedad semántica). Este módulo puede funcionar en paralelo con el módulo de análisis sintáctico o posteriormente. En cualquier caso, existen diferentes técnicas aplicables para resolver la ambigüedad: lógica de predicados, redes semánticas, grafos de dependencias conceptuales, etc.

Universitat d'Alacant
Universidad de Alicante

Según la Figura 1.2, los distintos módulos de un sistema de PLN necesitan fuentes externas tales como diccionarios, gramáticas u ontologías, que se adecúen al idioma o al dominio de los textos a tratar. Estas fuentes externas de conocimiento dotan al sistema de la información necesaria para poder establecer las reglas de la gramática, el dominio de aplicación de los textos, las relaciones entre palabras a partir de una jerarquía, etc.

Como veremos más adelante, todos los recursos léxicos utilizados por un sistema de PLN están basados en teorías psicolingüísticas adaptadas al idioma de estudio. En nuestro caso, se han utilizado diversos recursos léxicos externos para realizar la

tarea de desambiguación, que como indica la Figura 1.2 se realiza en el módulo de análisis semántico.

1.3 Objetivo de la Tesis

Dada la necesidad actual de tratamiento del lenguaje natural y descubierta la problemática de la ambigüedad intrínseca en el lenguaje humano, es necesario el tratamiento de forma eficaz de este problema.

Como se ha mencionado anteriormente, existen distintos tipos de ambigüedad: léxica, sintáctica y semántica. El tratamiento de los distintos tipos de ambigüedad requiere de la aplicación de técnicas de PLN específicas. Por tanto, es difícil intentar abarcar la resolución de todos los tipos de ambigüedad utilizando una misma técnica, y es por ello, que esta Tesis está centrada únicamente en la resolución de la ambigüedad semántica, que está vinculada a la aparición de palabras polisémicas en el lenguaje.

El objetivo principal de esta Tesis es desarrollar distintos métodos de resolución de la ambigüedad semántica basados en conocimiento. Cuando hablamos de métodos basados en conocimiento hacemos referencia a aquellos métodos que no necesitan de corpus extensos para poder funcionar, sino que utilizan recursos léxicos como diccionarios u ontologías para extraer las relaciones existentes entre palabras y construir estructuras de datos con información relevante que determine la similitud entre diferentes contextos, palabras, etc. Se ha demostrado que los resultados obtenidos con métodos basados en conocimiento son relativamente más bajos que los de otros métodos basados en técnicas de aprendizaje, pero la ventaja de los métodos basados en conocimiento radica en que su aplicación es muy útil en el caso de estudio de lenguas minoritarias, ya que encontrar corpus extensos de estas lenguas es muy complicado.

En esta Tesis se presentan varios métodos basados en conocimiento con distintos enfoques, aplicados a la desambiguación automática:

WSD DRelevant. Basado en la utilización de las categorías semánticas de WordNet Domains. El objetivo de este método es la determinación del sentido correcto de las palabras a partir del establecimiento del dominio contextual donde aparecen. El proceso comienza mediante la obtención de un nuevo recurso léxico (Dominios Relevantes) basado en la frecuencia de aparición de las palabras junto a diferentes dominios en WordNet. Este recurso es la base de todo el proceso de desambiguación centrado en medidas de frecuencia y co-ocurrencia.

WSD DLSA. Basado en la utilización de la técnica del Análisis de la Semántica Latente. Se construye un espacio semántico con las diferentes categorías de WordNet Domains. El objetivo de este método es establecer aquellas relaciones ocultas entre palabras que no pertenecen al mismo dominio o categoría semántica pero sí tienen alguna relación contextual.

WSD SenseDiscrim. Basado en la utilización de patrones y discriminadores de sentidos. El objetivo de este método es extraer ocurrencias de patrones léxicos a partir de diferentes contextos y determinar el sentido de las palabras presentes en dichos patrones utilizando la jerarquía de WordNet.

Todos los métodos descritos se basan en la premisa de que las palabras que aparecen en un mismo contexto tienden a estar relacionadas semánticamente. Es decir, es más probable que una palabra que aparezca en contextos similares tenga el mismo sentido, que esa misma palabra, en contextos dispares tenga el mismo significado.

Asimismo, la ambigüedad en el lenguaje no es un problema aislado en PLN, sino que afecta a diferentes áreas: traducción automática, extracción de información, recuperación de documentos, etc. Consecuentemente, en esta Tesis se han aplicado las técnicas de desambiguación automática propuestas, sobre algunas aplicaciones finales de PLN. Entre estas aplicaciones destacamos la detección de paráfrasis e implicación textual o la desambiguación y discriminación de entidades. El objetivo principal en este área, es demostrar que la inclusión de sistemas de resolución automática de la ambigüedad, es necesaria para la obtención de buenos

resultados. Además, el recurso léxico obtenido a partir de los dominios relevantes de WordNet, es una herramienta muy útil y que proporciona información beneficiosa para diferentes aplicaciones de PLN.

1.4 Organización de la Tesis

Esta Tesis se ha estructurado en siete capítulos:

- **Capítulo 1:** Introducción.
  Presentación de la problemática en el Procesamiento del Lenguaje Natural debida a la ambigüedad del lenguaje, objetivos y estructura de la Tesis.
- **Capítulo 2:** Estado del arte.
  Descripción de la evolución de los sistemas de desambiguación automática, clasificación y comparación de sistemas.
- **Capítulo 3:** Problemática en la evaluación de sistemas de WSD.
  Descripción de los distintos tipos de anotaciones utilizados para etiquetar los sentidos de las palabras, corpus, etc. Problemas de establecimiento de los sentidos correctos de las palabras y diferentes medidas de evaluación utilizadas.
- **Capítulo 4:** Recursos.
  Descripción de los recursos léxicos utilizados. Estudio de las ventajas e inconvenientes de cada recurso. Combinación de recursos con el objetivo de mejorar los resultados obtenidos.
- **Capítulo 5:** Métodos.
  Descripción de los métodos desarrollados. Aplicación a diversas tareas de PLN.
- **Capítulo 6:** Evaluación.
  Realización de diversos experimentos para la evaluación de los métodos. Comparativa de los resultados obtenidos con diferentes aproximaciones.
- **Capítulo 7:** Conclusiones y trabajos futuros.
  Aportación de esta Tesis al campo de la desambiguación automática y propuesta de trabajos futuros. Relación de publicaciones derivadas de la consecución de esta Tesis.

Universitat d'Alacant
Universidad de Alicante

Capítulo 2

Estado del arte

En este capítulo se presenta una breve introducción al problema de la ambigüedad semántica junto con las distintas aplicaciones en las que es necesario la aplicación de técnicas para resolver este tipo de ambigüedad. Además se establece la evolución de los diferentes sistemas de resolución de la ambigüedad semántica y su clasificación a partir de los recursos que utilizan.

2.1 Descripción del problema

Es muy común encontrar en cualquier idioma palabras con múltiples significados; por ejemplo, "flojo" puede significar algo que está poco apretado o alguien que es un cobarde. El significado particular de una palabra viene determinado por el contexto que la rodea y en muchas ocasiones por la situación en que se emplea. Si por ejemplo, decimos: "Te has dejado los cordones de los zapatos demasiado *flojos*", en este caso, sabemos exactamente el significado asociado a esta palabra.

El procedimiento para decidir los significados de las palabras a partir del contexto que las rodea se conoce como "desambiguación" o *"Word Sense Disambiguation"* (WSD).

En procesamiento del lenguaje natural las investigaciones en WSD han existido desde la aparición de este área de investigación.

Es más, WSD se ha considerado como una tarea completamente distinta a otras dirigidas al usuario final, como por ejemplo, Traducción Automática. De hecho, para obtener un buen sistema de Traducción Automática es necesario resolver el problema de la ambigüedad, y poder proporcionar de esta forma, una buena comprensión del lenguaje.

A continuación se van a describir algunas de los métodos más conocidos para resolver el problema de WSD. Además, se mostrarán los avances más recientes en este campo, dentro de una de las competiciones más importantes: SENSEVAL (Kilgarriff (1998b)).

2.2 Aplicaciones de WSD

En traducción automática la desambiguación es uno de los principales problemas que necesitan tratamiento. Aunque este es uno de los principales usos de WSD también deben considerarse otras muchas aplicaciones de PLN que necesitan resolver este problema. Recientemente han aparecido nuevas áreas de conocimiento y el tratamiento automático para la resolución de la ambigüedad es muy necesario. Entre estas nuevas áreas de conocimiento encontramos por ejemplo la bioinformática y la Web Semántica.

Universitat d'Alacant
Universidad de Alicante

A continuación vamos a ver qué necesidades tienen las diferentes tareas de PLN y de qué forma se aplica la desambiguación automática para su correcto funcionamiento:

Traducción automática (TA). WSD es necesaria para la selección de la correcta traducción de palabras que pueden tener distintas acepciones según el sentido asignado. Por ejemplo, en un texto que necesite ser traducido del inglés al francés podemos encontrar la palabra "*change*" que puede ser traducida como "*changement*" (transformación) o "*monnaie*" (dinero suelto). En TA, los sentidos se representan a menudo como palabras en el lenguaje de traducción destino. Sin embargo, muchos modelos de TA no utilizan WSD explícitamente. En muchos casos el vocabulario es pre-desambiguado para un dominio determinado, se desarrollan

reglas hechas a mano o WSD está almacenado en un modelo de traducción estadístico (Brown et al. (1991)).

Recuperación de Información (RI). En este caso, la ambigüedad debe ser resuelta en algunas cuestiones. Por ejemplo, si tenemos una cuestión en la que aparece la palabra "*depression*", el sistema de RI podría devolver documentos que hablan sobre enfermedades, el tiempo o economía. Un problema similar aparece asociado a nombres propios tales como "*Raleigh*" (bicicleta, persona, ciudad, etc). Los sistemas actuales de RI no utilizan explícitamente WSD, únicamente se basan en que el usuario dé el suficiente contexto en la pregunta para extraer únicamente los documentos relevantes al sentido correcto (ej "*tropical depression*"). Experimentos recientes sugieren que un sistema fiable de RI necesitaría al menos un 90 % de precisión en WSD (Sanderson (1994)). Recientemente, se ha comprobado que WSD mejora la *"cross-lingual IR"* y la clasificación de documentos (Vossen et al. (2006), Bloehdorn y Hotho (2004), Clough y Stevenson (2004)). Otras aplicaciones relacionadas con clasificación de documentos y *"cross-lingual IR"* incluyen recomendación de noticias, alertas y posicionamiento automático de publicidad.

Extracción de Información (EI) y Minería de Textos. WSD es necesaria para el correcto análisis de los textos en muchas aplicaciones. Por ejemplo, las investigaciones en bioinformática requieren establecer las relaciones entre genes y productos genéticos para ser catalogados a través de la amplia literatura científica. Sin embargo, los genes y sus proteínas a menudo tienen el mismo nombre. De forma más general, la Web Semántica requiere anotación automática de documentos de acuerdo a una ontología de referencia: todas las referencias textuales deben estar asignadas a conceptos y estructuras de eventos en la ontología (Dill et al. (2003)). La clasificación de entidades, la determinación de co-referencias y la expansión de acrónimos (MG como magnesio o miligramos) puede también clasificarse como un problema de WSD para nombres propios. Actualmente, WSD está empezando a aplicarse en estas áreas.

Lexicografía. La lexicografía moderna está basada en corpus, por tanto, WSD y lexicografía pueden trabajar conjuntamente,

de forma que WSD proporcione grupos e indicadores contextuales significativos de sentidos a los lexicógrafos, los cuales, producirán mejores inventarios de sentidos y de corpus anotados para WSD.

A pesar de este amplio rango de aplicaciones donde WSD muestra un gran potencial para ser de utilidad, aún no se ha demostrado que proporcione mejoras significativas en este tipo de aplicaciones. Existen varios resultados aislados que muestran muy pocas mejoras y en algunos casos puede perjudicar el rendimiento como muestra un experimento realizado sobre RI (Sanderson (1994)). Existen varias posibles razones para esto. En primer lugar, el dominio de una aplicación a menudo restringe el número de sentidos que una palabra puede tener (por ejemplo, no esperaríamos tener "banco" con el sentido "Conjunto de peces que van juntos en gran número" en un documento que hable sobre finanzas), por tanto, los lexicones pueden construirse adaptados al dominio. En segundo lugar, WSD todavía no está lo suficientemente desarrollada como para mostrar un efecto significativo. En tercer lugar, tratar la WSD como un módulo específico, significa que no puede integrarse apropiadamente dentro de una aplicación particular o ser entrenada dentro de un dominio específico. Muchas aplicaciones como TA, no tienen un lugar para un módulo de WSD, por tanto, o la aplicación o el módulo de WSD deben ser rediseñados.

A pesar de todo ello, queda patente que las aplicaciones requieren de WSD de alguna forma. Por ejemplo, en RI, una cuestión de dos palabras puede desambiguarse implícitamente, debido a que ambas palabras se utilizan juntas en los textos con su correspondiente sentido asociado (por ejemplo, "*tropical depression*"). El trabajo en WSD puede servir para explorar y remarcar las características particulares que proporcionen mejores resultados para una desambiguación más precisa.

2.3 Sistemas iniciales: el comienzo

Toda área de investigación tiene sus comienzos y por supuesto en WSD también existieron los primeros sistemas desarrollados alrededor de los años 70 y 80.

Uno de los primeros sistemas que trataron de resolver el problema de la ambigüedad léxica fue el creado por (Wilks (1972)). En este caso, se utilizaron restricciones de selección organizadas jerárquicamente junto con representaciones semánticas complejas denominadas fórmulas.

En este sistema, la hipótesis era que para cada uno de los diferentes sentidos de una palabra existía una fórmula asociada. El sistema incluía una jerarquía de ocho características semánticas: HUMAN, WANT, ABSTRACT, etc. Las fórmulas para los adjetivos contenían las preferencias semánticas de los nombres a los que podían acompañar, al igual que las fórmulas para los verbos que contenían las preferencias de los nombres con los que estaban relacionados. La polisemia venía determinada al asignar más de una fórmula a una misma palabra. La forma de obtener los sentidos de cada palabra era obteniendo una fórmula para cada una de las palabras, que maximizara el número de preferencias para una frase determinada. De esta forma, se podía establecer el sentido de todas las palabras de una frase simultáneamente. Finalmente, la evaluación del sistema se realizó sobre textos obtenidos a partir de artículos de periódicos.

Otro sistema realizado en 1980 (S. (1980)) tenía como hipótesis que el conocimiento humano sobre el lenguaje se debe principalmente al conocimiento sobre palabras más que al conocimiento de reglas (Small y Rieger (1982)). Sin embargo, este punto de vista es poco convencional y no se han realizado estudios que sostengan esta teoría psicolingüística.

Universitat d'Alacant
Universidad de Alicante

2.4 Clasificación de sistemas en WSD

Desde los primeros sistemas de WSD hasta la actualidad han surgido nuevas propuestas y distintos enfoques para resolver este problema. Una forma muy extendida de clasificar los sistemas de WSD es basándose en la principal fuente de conocimiento utilizada para establecer los diferentes sentidos. En primer lugar, tenemos los métodos que utilizan diccionarios, tesauros y bases de conocimiento léxicas, sin utilizar ningún corpus (etiquetado o

no). Este tipo de métodos son los denominados "*dictionary-based*" o "*knowledge-based*". Por otra parte, tenemos aquellos métodos que evitan casi completamente la información externa y trabajan directamente con corpus sin etiquetar, son los denominados métodos no supervisados. Dentro de esta categoría encontramos los métodos que utilizan "*word-aligned corpora*" para acumular "*cross-linguistic evidence*" para discriminación de sentidos. Finalmente, tenemos los sistemas supervisados y semi-supervisados, estos métodos utilizan corpus anotados semánticamente como entrenamiento, o como fuente de datos en un sistema de "*bootstrapping*".

Casi todas las aproximaciones de aprendizaje supervisado se han aplicado a WSD, incluyendo algoritmos agregativos y discriminativos y técnicas asociativas tales como selección de características, optimización de parámetros, etc.

Los métodos no supervisados tienen la ventaja de evitar el cuello de botella existente en la adquisición de nuevo conocimiento (anotación manual) (Boguraev y Briscoe (1989), Pustejovsky (1991)) y han obtenido buenos resultados (Schütze (1998)). Estos métodos son capaces de inducir sentidos de palabras, a partir de textos de entrenamiento, agrupando (mediante clusters) ocurrencias de palabras y clasificando entonces nuevas ocurrencias en los clusters inducidos.

Las propuestas basadas en conocimiento de los años 1970 y 1980 están todavía en proceso de investigación. Las principales técnicas utilizan restricciones de selección, el solapamiento de textos y medidas de similitud semántica. Actualmente, la tendencia es hacer una inferencia semántica general utilizando bases de conocimiento, obteniendo como resultado una desambiguación.

En la Tabla 2.1 tenemos un resumen de las distintas aproximaciones a WSD.

2.5 Métodos basados en conocimiento

En esta categoría encontramos diferentes algoritmos para la etiquetación automática de sentidos. Normalmente, el rendimien-

Métodos	Procedimiento
Basados en conocimiento	Creación de reglas de desambiguación Restricciones de selección (o preferencias), utilizadas para filtrar sentidos incongruentes Comparación de las definiciones de los diccionarios con el contexto (método de Lesk) Selección del sentido más similar al contexto, utilizando medidas de similitud semántica Un sentido por discurso y otras heurísticas
Basados en corpus no supervisados	Métodos no supervisados que clasifican palabras o contextos en diferentes clusters, obteniendo los diferentes sentidos Utilización de corpus paralelos para inferir sentidos entre diferentes idiomas
Basados en corpus supervisados	Aprendizaje automático supervisado, utilizando corpus de entrenamiento etiquetados manualmente Métodos de bootstrapping (semi-supervisados)
Métodos híbridos	Utilización de técnicas de clustering no supervisadas combinadas con métodos basados en conocimiento Utilización de métodos basados en conocimiento para buscar ejemplos que sirvan de entrenamiento en métodos supervisados Utilización de corpus paralelos combinados con métodos basados en conocimiento

Tabla 2.1. Clasificación de métodos de WSD

to de estos métodos basados en conocimiento, es menor en comparación con los métodos basados en corpus. Pero con la salvedad de que los métodos basados en conocimiento tienen una amplia cobertura ya que pueden aplicarse a cualquier tipo de texto en comparación con los basados en corpus que sólo se pueden aplicar a aquellas palabras de las que se dispone de corpus anotados.

A continuación vamos a enumerar diferentes técnicas utilizadas por los métodos basados en conocimiento, aplicables sobre cualquier base de conocimiento léxica que defina sentidos de palabras y relaciones entre ellas. La base de conocimiento léxica más utilizada es WordNet (Miller (1995)). Vamos a describir 4 tipos diferentes de métodos basados en conocimiento:

1. El algoritmo de Lesk, en el cual, los sentidos de las palabras de un contexto se identifican basándose en una medida de solapamiento contextual entre las definiciones de un diccionario.

2. Medidas de similitud semántica extraídas a través de redes semánticas. Esta categoría incluye métodos que tratan de encontrar la distancia semántica existente entre diferentes conceptos. Dependiendo del tamaño del contexto estas medidas se dividen en dos grandes categorías:
   - Métodos aplicables a un contexto local, donde las medidas de similitud semántica se utilizan para desambiguar palabras conectadas por relaciones sintácticas o por su localización.
   - Métodos aplicables a contextos globales, donde las cadenas léxicas son derivadas basándose en medidas de similitud semántica (una cadena léxica es un hilo de significado extraído a través del texto total).
3. Preferencias de selección adquiridas de forma automática o semi-automática, como una forma de restringir los posibles sentidos de una palabra, basados en la relación que ésta tiene con otras palabras en el contexto.
4. Métodos heurísticos, que consisten en reglas que pueden asignar un sentido a ciertas categorías de palabras, incluyendo:
   - El sentido más frecuente.
   - Un sentido por colocación.
   - Un sentido por discurso.

Estos cuatro tipos de métodos se van a tratar en detalle en las siguientes secciones.

Universitat d'Alacant
Universidad de Alicante

2.5.1 Algoritmo de Lesk

El algoritmo de Lesk (Lesk (1986)) es uno de los primeros algoritmos desarrollados para la desambiguación semántica de todas las palabras en cualquier texto. El único recurso requerido por el algoritmo es un conjunto de entradas de un diccionario, una por cada posible sentido y conocimiento sobre el contexto inmediato donde se desarrolla la desambiguación.

Aunque este algoritmo se considera un método basado en diccionarios, también es el punto de partida para los algoritmos basados en corpus. Casi todos los algoritmos supervisados se basan de alguna forma en solapamiento contextual, midiendo ese solapa-

miento entre el contexto de una palabra ambigua y los contextos específicos para cada uno de los sentidos de esa palabra.

La principal idea de este algoritmo es desambiguar palabras encontrando el solapamiento entre las definiciones de sus sentidos. En otras palabras, dadas dos palabras, W_1 y W_2, cada una con sus respectivos sentidos N_{w_1} y N_{w_2} definidos en un diccionario, para cada par de posibles sentidos W_1^i y W_2^j, $i = 1..N_{w_1}$, $j = 1..N_{w_2}$, primero se determina el solapamiento con las correspondientes definiciones contando el número de palabras que tienen en común. A continuación, el par de sentidos con el mayor solapamiento es seleccionado y entonces se le asigna un sentido a cada palabra del par inicial. En la Tabla 2.2 se muestran los principales puntos de este algoritmo.

(1) Para cada sentido i de W_1
(2) Para cada sentido j de W_2
(3) Calcular el solapamiento(i, j), el número de palabras en común entre las definiciones del sentido i y el sentido j
(4) Encontrar i y j tales que el solapamiento(i, j) sea el máximo
(5) Asignar el sentido i a W_1 y el sentido j a W_2

Tabla 2.2. Algoritmo de Lesk

Un ejemplo representativo de este algoritmo sería el siguiente:

Consideremos que queremos desambiguar las palabras "*pine*" y "*cone*", mediante el par de palabras "*pine cone*". El diccionario Oxford Advanced Learner's define cuatro sentidos para "*pine*" y tres sentidos para "*cone*", tal y como muestra la Tabla 2.3.

En la Tabla 2.4 podemos ver el solapamiento existente entre cada sentido de "*pine*" y cada sentido de "*cone*".

La primera definición de "*pine*" y la tercera de "*cone*" tienen el máximo solapamiento entre todas las posibles combinaciones de sentidos, con dos palabras en común: "*evergreen*" y "*tree*", por lo tanto, estos son los sentidos seleccionados por el algoritmo de Lesk.

Este algoritmo fue evaluado sobre un conjunto de pares de palabras ambiguas manualmente anotados, utilizando el diccionario

pine
1* seven kinds of evergreen tree with needle-shaped leaves
2 pine
3 waste away through sorrow or illness
4 pine for something, pine to do something
cone
1 solid body which narrows to a point
2 something of this shape, whether solid or hollow
3* fruit of certain evergreen trees (fir, pine)

Tabla 2.3. Definiciones para "*pine*" y "*cone*"

$Pine\#1 \bigcap Cone\#1 = 0$
$Pine\#2 \bigcap Cone\#1 = 0$
$Pine\#3 \bigcap Cone\#1 = 0$
$Pine\#4 \bigcap Cone\#1 = 0$
$Pine\#1 \bigcap Cone\#2 = 0$
$Pine\#2 \bigcap Cone\#2 = 0$
$Pine\#3 \bigcap Cone\#2 = 1$
$Pine\#4 \bigcap Cone\#2 = 0$
$Pine\#1 \bigcap Cone\#3 = 2$
$Pine\#2 \bigcap Cone\#3 = 1$
$Pine\#3 \bigcap Cone\#3 = 0$
$Pine\#4 \bigcap Cone\#3 = 1$

Tabla 2.4. Solapamiento entre "*pine*" y "*cone*"

Universitat d'Alacant
Universidad de Alicante

Oxford Advanced Learner's, obteniendo una precisión entre 50 y 70 % (Lesk (1986)).

2.5.2 Variaciones del algoritmo de Lesk

Desde el planteamiento inicial del algoritmo de Lesk en 1986 se han propuesto varias variantes del algoritmo, incluyendo:

- Versiones del algoritmo que tratan de resolver el problema de la explosión combinatoria cuando se consideran más de dos palabras.
- Versiones del algoritmo donde cada palabra de un contexto determinado es desambiguada individualmente midiendo el solapamiento entre cada definición del diccionario y el contexto en el cual aparece.

- Alternativas donde el espacio semántico de una palabra es aumentado con definiciones de palabras relacionadas semánticamente.

2.5.2.1 Simulated Annealing.

Una de las principales desventajas del algoritmo de Lesk inicial, es que conlleva una explosión combinatoria cuando se trata de aplicar a la desambiguación de más de dos palabras. Por ejemplo, podemos considerar el texto *"I saw a man who is 98 years old and can still walk and tell jokes"*, con nueve palabras a desambiguar, cada una de ellas con sus correspondientes sentidos: "*see*(26)", "*man*(11)", "*year*(4)", "*old*(8)", "*can*(5)", "*still*(4)", "*walk*(10)", "*tell*(8)", "*joke*(3)". Un total de 43929600 combinaciones de sentidos pueden ser posibles para este texto, por lo tanto, el algoritmo de Lesk original no es una aproximación óptima para este problema.

Una solución posible sería utilizar el algoritmo *"simulated annealing"*, propuesto por Cowie et al. (Cowie et al. (1992)). En esta propuesta, se define una función E que refleja las combinaciones de sentidos en un texto, y cuyo valor mínimo se corresponde con la selección de los sentidos correctos. Para una combinación dada de sentidos, se extraen todas las definiciones correspondientes de un diccionario y cada palabra que aparece en una de estas definiciones recibe un valor igual a su número de ocurrencias. Uniendo todos estos valores se obtiene la "redundancia" del texto. La función E se define entonces como la inversa de la redundancia, siendo el objetivo final encontrar la combinación de sentidos que minimice esta función. Para este propósito, se determina una combinación inicial de sentidos (por ejemplo, se recogen los sentidos más frecuentes para cada palabra), y entonces se realizan varias iteraciones, donde el sentido de una palabra aleatoria en el texto se reemplaza con otro sentido distinto, y la nueva selección se considera correcta únicamente si reduce el valor de la función E. Las iteraciones terminan cuando no existe ningún cambio en la configuración de los sentidos. La evaluación de este método sobre 50 frases de ejemplo consiguió un 47 % de precisión a nivel

de sentidos y un 72 % de precisión a nivel de homógrafos. Este método fue re-implementado por Stevenson y Wilks (Stevenson y Wilks (2001)) obteniendo un valor similar de precisión, en torno a un 65,24 % en un corpus etiquetado con los sentidos del *Longman Dictionary of Contemporary English*[1] (LDOCE).

2.5.2.2 Algoritmo de Lesk simplificado.

Otra versión del algoritmo de Lesk, que también trata de resolver el problema de la explosión combinatoria, es una variación simplificada que utiliza un proceso separado de desambiguación para cada palabra ambigua del texto de entrada. En este algoritmo simplificado, el sentido correcto de cada palabra en un texto, se determina individualmente, encontrando el sentido que lleva al máximo solapamiento entre las definiciones del diccionario y el contexto actual. Esta aproximación toma cada palabra de forma individual, sin tener en cuenta el sentido de las otras palabras que aparecen junto a ella. En la tabla 2.5 se muestran los principales pasos de este algoritmo simplificado.

(1) Para cada sentido i de W
(2) Determinar el Solapamiento(i), el número de palabras en común entre la definición del sentido i y el contexto donde aparece la palabra
(3) Encontrar el sentido i con el máximo Solapamiento(i)
(4) Asignar el sentido i a W

Tabla 2.5. Algoritmo simplificado de Lesk

Un estudio realizado por (Vasilescu et al. (2004)) ha demostrado que el algoritmo simplificado de Lesk mejora la definición original del algoritmo en términos de precisión y eficiencia. Su evaluación se realizó utilizando los datos de la tarea *"English all-words"* de SENSEVAL-2, obteniendo un 58 % de precisión con el algoritmo simplificado, por encima del 42 % obtenido por el algoritmo original.

[1] `http://www.longman.com/ldoce`

Otra versión del algoritmo de Lesk utiliza corpus anotados para resolver la ambigüedad de una palabra determinada. En este caso, esta versión tiene la capacidad de aumentar el contexto de una palabra con ejemplos adicionales etiquetados. Por lo tanto, el sentido seleccionado para la aparición de una palabra en un nuevo contexto, será aquel que tenga mayor solapamiento con algún contexto pre-etiquetado anteriormente.

En la tabla 2.6 se muestran los pasos del algoritmo de Lesk basado en corpus, suponiendo que se dispone de ejemplos etiquetados de la palabra a desambiguar.

(1) Para sentido i de W
(2) Se establece Peso(i) a 0
(3) Para cada palabra única w en contexto de W
(4) si w aparece en los ejemplos etiquetados o en la definición del diccionario del sentido i
(5) Seleccionar el sentido i con el mayor Peso(i)

Tabla 2.6. Algoritmo de Lesk basado en corpus

El *Peso* de una palabra se define usando una medida extraída de los métodos de Recuperación de Información: $Peso(w)$ es la inversa de la frecuencia en documentos (idf) de la palabra sobre los ejemplos y las definiciones del diccionario. El idf de una palabra es $\log(p(w))$, donde $p(w)$ se define como la fracción de ejemplos y definiciones que incluyen la palabra w.

Esta nueva aproximación ha conseguido los mejores resultados en comparación con los métodos de aprendizaje supervisado. En SENSEVAL-1 (Kilgarriff y Rosenzweig (2000)) la aproximación del algoritmo de Lesk basada en corpus obtuvo un 69,1 % de precisión comparado con el 56,6 % obtenido utilizando la heurística del sentido más frecuente. En SENSEVAL-2 (Kilgariff (2001)) el algoritmo de Lesk consiguió resultados similares: 51,2 % precisión comparado con el 64,2 % conseguido por el mejor sistema supervisado.

2.5.2.3 Espacios semánticos aumentados.

Otra variante del algoritmo de Lesk es la propuesta por Banerjee y Pedersen (Banerjee y Pedersen (2002)) denominada *Algoritmo de Lesk Adaptado*. En esta propuesta se utilizan junto con las definiciones de la palabra ambigua, las definiciones de palabras relacionas. En este caso, se utiliza una función similar a la empleada por (Cowie et al. (1992)) para determinar el valor para cada combinación posible de sentidos en un texto, y tratar de identificar la combinación que lleva al máximo valor.

En esta aproximación se tienen en cuenta conceptos relacionados con la palabra ambigua utilizando la jerarquía de WordNet: hiperónimos, hipónimos, holónimos, merónimos, tropónimos. Se utilizan relaciones de atributos y sus correspondientes definiciones, para construir un contexto más amplio a partir de las definiciones semánticas.

2.5.3 Similitud semántica

En la técnica basada en similitudes semánticas, la premisa inicial es que las palabras de un texto deben relacionarse según sus sentidos para obtener un discurso coherente (Halliday y Hasan (1976)). Esta premisa es una propiedad natural del lenguaje humano y al mismo tiempo la base para el desarrollo de los sistemas de desambiguación automáticos. Se puede afirmar, por tanto, que las palabras que comparten un contexto similar están normalmente relacionadas y por consiguiente, se pueden seleccionar sus sentidos a partir de la distancia semántica (Rada et al. (1989)).

Esta premisa se restringe a un pequeño grupo de palabras extraídas del contexto más cercano a la palabra ambigua o a las palabras relacionadas sintácticamente con la palabra ambigua. Este tipo de métodos extrae el contexto local y no introduce información contextual adicional obtenida a partir de una ventana de cierto tamaño.

Existen otros métodos que utilizan un contexto global y tratan de construir hilos de conocimiento a través del texto completo, utilizando para ello ventanas centradas en la palabra ambigua.

Al igual que sucedía con el algoritmo de Lesk, estos métodos sufren de un gran coste computacional cuando tratan más de dos palabras. Para resolver este problema se pueden aplicar las mismas soluciones propuestas para el algoritmo de Lesk, como por ejemplo el algoritmo propuesto por (Agirre y Rigau (1996)).

2.5.3.1 Medidas de similitud semántica.

Existen diferentes medidas de similitud que tratan de cuantificar el grado en que dos palabras están relacionadas semánticamente. Muchas de estas medidas se basan en redes semánticas y siguen la metodología original propuesta por (Rada et al. (1989)).

A continuación se muestran una serie de medidas de similitud aplicadas sobre la jerarquía de WordNet. La mayoría de estas medidas toman como entrada un par de conceptos y devuelven un valor que indica el grado de similitud entre ambas palabras.

1. (Leacock et al. (1998)). Esta medida determina el camino mínimo entre las dos palabras de entrada. Este valor se normaliza atendiendo a la profundidad de la taxonomía. En la Ecuación (2.1) $Camino(C_1, C_2)$ representa la longitud del camino que conecta los dos conceptos (es decir, el número de arcos en la red semántica que son atravesados para llegar de C_1 a C_2, y D es la profundidad total de la taxonomía.

$$Similitud(C_1, C_2) = -log\left(\frac{Camino(C_1, C_2)}{2D}\right) \tag{2.1}$$

2. (Hirst y St-Onge (1998)). Añaden a la medida de similitud la dirección de los enlaces que forman el camino. Además de la longitud, el camino no debería "cambiar de dirección frecuentemente". En la Ecuación (2.2) C y k son constantes, el $Camino$ se define como en la ecuación (2.1) y d representa el número de cambios de dirección.

$$Similitud(C_1, C_2) = C - Camino(C_1, C_2) - kd \tag{2.2}$$

3. (Resnik (1995b)). Define el término de contenido de información, que es una medida de la especificación de un concepto determinado, y está definida en base a su probabilidad de ocurrencia en un corpus extenso.

$$IC(C) = -log(P(C)) \tag{2.3}$$

Dado un corpus, $P(C)$ es la probabilidad de encontrar una instancia de tipo C. El valor para $P(C)$ es mayor para conceptos listados en la parte superior de la jerarquía y llega a su máximo valor para el concepto que se encuentra en la cima (si la jerarquía tiene tiene una única cima, entonces el valor para este concepto es 1).
Resnik define una medida de similitud semántica entre dos palabras utilizando el *"Lowest Common Subsumer"* (LCS). El LCS es el primer concepto de la red semántica que contiene a las dos palabras, es decir, el primer nodo común para el cual existe un camino desde la palabra W_1 y la palabra W_2. En la Ecuación (2.4) se muestra esta medida.

$$Similitud(C_1, C_2) = IC(LCS(C_1, C_2)) \tag{2.4}$$

4. (Jiang y Conrath (1997)) presentan una alternativa a la medida de Resnik utilizando la diferencia existente en el contenido de información de los dos conceptos para indicar su similitud. Como muestra la Ecuación (2.5).

$$\begin{aligned} Similitud(C_1, C_2) = {} & 2 \,\mathrm{x}\, IC(LCS(C_1, C_2)) \\ & - (IC(C_1) + IC(C_2)) \end{aligned} \tag{2.5}$$

Además de esta aproximación (Lin (1998b)) desarrolla otra fórmula que combina la información de LCS con la información de los conceptos involucrados, según la ecuación (2.6).

$$Similitud(C_1, C_2) = \frac{2 \,\mathrm{x}\, IC(LCS(C_1, C_2))}{IC(C_1) + IC(C_2)} \tag{2.6}$$

Universidad de Alicante

5. (Mihalcea y Moldovan (1999)) introducen una nueva fórmula para medir la similitud semántica entre jerarquías independientes, incluyendo jerarquías para diferentes categorías léxicas. Todas las medidas de similitud comentadas anteriormente sólo se pueden aplicar a conceptos que están explícitamente conectados por algún arco en la red semántica. Con esta nueva medida Mihalcea y Moldovan crean caminos virtuales entre diferentes jerarquías a través de las definiciones de las glosas en WordNet. En la Ecuación (2.7) $|CD_{12}|$ es el número de palabras comunes a las definiciones en la jerarquía de C_1 y C_2. $descendientes(C_2)$ es el número de conceptos en la jerarquía de C_2. Y W_k es un peso asociado con cada concepto determinado como la profundidad del concepto dentro de la jerarquía.

$$Similitud(C_1, C_2) = \frac{\sum_{k=1}^{|CD_{12}|} W_k}{\log(descendientes(C_2))} \tag{2.7}$$

   Esta medida de similitud funciona bastante bien para la desambiguación de nombres y verbos conectados por una relación sintáctica (por ejemplo, verbo-objeto).
6. (Agirre y Rigau (1996)) introducen la noción de densidad conceptual, definida como el solapamiento entre la jerarquía semántica enraizada por un concepto C, y las palabras en el contexto de C. En la Ecuación (2.8), m es el número total de sentidos en el contexto de C encontrados en la jerarquía cuya raíz es C, y $descendientes(C)$ representa el total del número de conceptos en la jerarquía enraizada por C. W_k es un peso asociado con cada concepto en la jerarquía ($nhyp$ es el número de hipónimos para un nodo determinado en la jerarquía, y el valor óptimo para α fue determinado empíricamente a 0,20).

$$DC(C) = \frac{\sum_{k=0}^{m} W_k}{descendientes(C)},\ donde\ W_k = nhyp^{k^\alpha} \tag{2.8}$$

   Para identificar el sentido de una palabra en un determinado contexto, la fórmula de la densidad conceptual se aplica a todos los posibles sentidos de la palabra, escogiendo el sentido cuya densidad conceptual sea mayor.

2.5.3.2 Similitud semántica en un contexto local.

La aplicación de las medidas de similitud mostradas anteriormente sobre cualquier texto, no es un proceso sencillo. Generalmente, en un texto encontramos más de dos palabras ambiguas, por tanto, debemos tratar con conjuntos de palabras ambiguas donde la distancia de una palabra al resto de palabras en el contexto influye sobre el sentido adoptado.

Los trabajos desarrollados dentro de este área utilizan un contexto local restringiendo así el número de palabras ambiguas dentro del mismo contexto. (Patwardhan et al. (2003)) aplican la primera medida de similitud de la lista anterior para decidir el sentido correcto de las 1723 instancias de nombres de la tarea *"English Lexical Sample"* de SENSEVAL-2. En este estudio, se utiliza un valor acumulativo añadiendo las distancias semánticas de la palabra a desambiguar junto con las palabras vecinas (una palabra a la izquierda y una palabra a la derecha). El sentido elegido es aquel cuyo valor acumulado es mayor. Tras el proceso de evaluación se determinó que entre las medidas de similitud propuestas, (Jiang y Conrath (1997)) alcanzaban la mejor puntuación y (Hirst y St-Onge (1998)) proporcionaban el mejor funcionamiento a través de varias palabras.

Las dependencias sintácticas son otra posible restricción a considerar. En este caso, (Stetina et al. (1998)) proponen un método basado en las relaciones sintácticas de palabras y una medida muy simple que define que dos palabras son similares si pertenecen al mismo synset de WordNet.

En (Montoyo (2002)) se propone la identificación del sentido correcto de las palabras a través del algoritmo de *Marcas de Especifidad*. Este algoritmo utiliza la taxonomía de nombres de WordNet, sus relaciones de hiponimia e hiperonimia, para desambiguar palabras dentro de un contexto local (oración). La hipótesis en la que se basa este algoritmo es que las palabras que aparecen en un mismo contexto tienen sus sentidos relacionados entre sí, y por tanto, puede existir un concepto dentro de la red semántica que relacione todas las palabras del contexto. Este concepto superior es la denominada *Marca de Especifidad* (ME). El proce-

so de desambiguación es el siguiente: a través de la jerarquía de WordNet y de las palabras del contexto, se obtiene el conjunto de hiperónimos/hipónimos que comparten las palabras. Usando esta información se trata de determinar el concepto superior (ME) que engloba el mayor número de palabras del contexto con sus respectivos sentidos. Si como resultado para la ME inicial aún existen palabras ambiguas en el contexto, se va descendiendo por la jerarquía obteniendo nuevas ME, de forma que se seleccionará aquella ME que maximice el número de palabras del contexto no ambiguas. La Figura 2.1 muestra su funcionamiento.

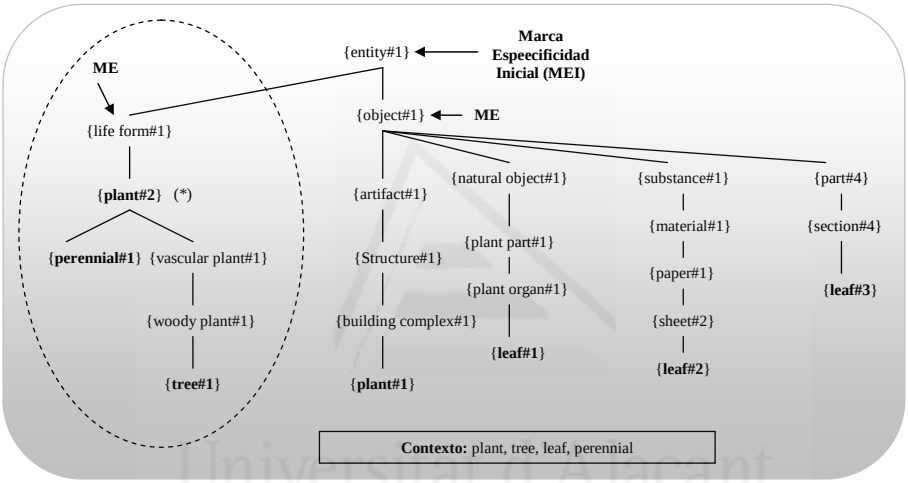


Figura 2.1. Algoritmo Marcas de Especifidad

Universitat d'Alacant
Universidad de Alicante

2.5.3.3 Similitud semántica en un contexto global.

Las cadenas léxicas son una de las estructuras de conocimiento más comunes. Una cadena léxica es una secuencia de palabras relacionadas semánticamente, lo cual crea un contexto y contribuye a la continuidad del conocimiento y de la coherencia de un discurso (Halliday y Hasan (1976)). Estas estructuras han sido consideradas muy útiles en diferentes tareas de procesamiento del lenguaje natural, incluyendo resumen automático, categorización de textos y desambiguación automática. Las cadenas léxicas se

extraen independientemente de la estructura gramatical del texto y pueden abarcar grandes distancias dentro del texto.

Un algoritmo genérico de creación de cadenas léxicas se divide en tres pasos (Figura 2.2):

1. Seleccionar las palabras candidatas del texto. Éstas son palabras a partir de las cuales podemos establecer similitudes semánticas y por tanto, la mayor parte del tiempo pertenecen a la misma categoría léxica.
2. Para cada una de estas palabras candidatas, y para cada sentido, se busca una cadena que reciba el sentido de la palabra candidata, basándose en una medida de similitud entre los conceptos que ya están en la cadena y el sentido de la palabra candidata.
3. Si esa cadena se encuentra, se inserta la palabra dentro de la cadena, en otro caso, se crea una nueva cadena.

Todas las cadenas que superan un cierto umbral son seleccionadas.

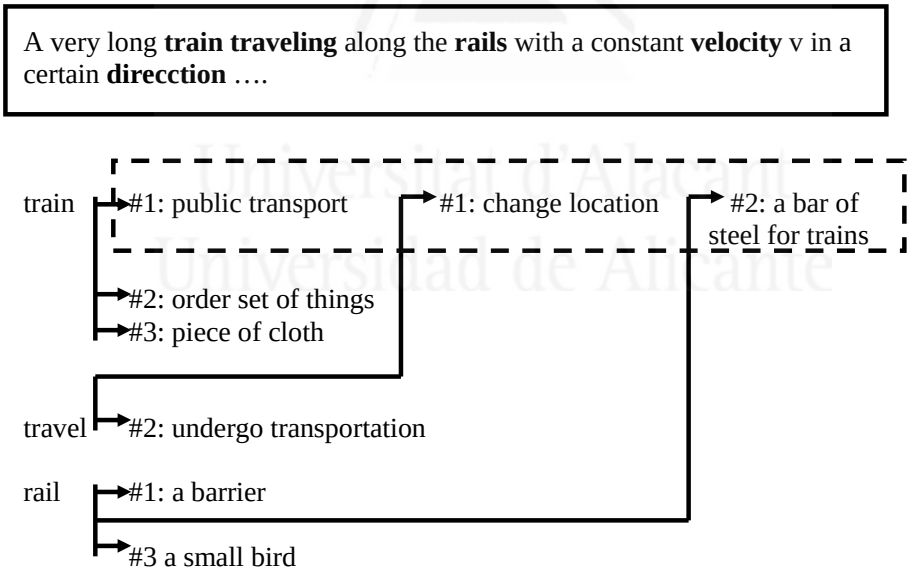


Figura 2.2. Cadenas léxicas en un contexto global

2.5.4 Preferencias de selección

Algunos de los algoritmos creados inicialmente para WSD se basan en preferencias de selección como una forma de restringir los posibles sentidos de una palabra en un contexto determinado.

Las preferencias de selección capturan información sobre las posibles relaciones entre diferentes categorías de palabras según el propio sentido común. Por ejemplo, COMER-COMIDA o BEBER-LÍQUIDOS, son muestras de tales restricciones semánticas, las cuales pueden ser utilizadas para desechar sentidos incorrectos y seleccionar sólo aquellos sentidos que se corresponden con los sentidos obtenidos siguiendo las reglas. Por ejemplo, dada la frase *"Mary drunk burgundy"*, el sentido para *"burgundy"* que lo define como un color, no tiene cabida en este contexto porque el verbo *"drink"* requiere de un líquido como objeto directo.

2.5.4.1 Adquisición de preferencias de selección.

A pesar de que las preferencias de selección son intuitivas, es muy difícil ponerlas en práctica para resolver el problema de WSD. Supongamos, por ejemplo, que queremos obtener un corpus anotado semánticamente. En este caso, el principal problema es la relación circular entre las preferencias de selección y WSD, ya que, aprender restricciones semánticas requiere conocimiento sobre los sentidos involucrados en una relación y viceversa. En (Brockmann y Lapata (2003)) se realizó un estudio sobre métodos que utilizan preferencias de selección junto con una evaluación de los resultados obtenidos de forma automática frente a los resultados anotados por un ser humano.

Otra alternativa para obtener preferencias de selección es partir de corpus no anotados semánticamente. De esta forma se pueden utilizar técnicas estadísticas para establecer relaciones entre palabras:

Contador de frecuencia:

$$Cont\_frec(W_1, W_2, R) \tag{2.9}$$

Se determina cuántas veces co-ocurre la palabra W_1 con la palabra W_2 mediante la relación R.

Probabilidad condicional:

$$P(W_1|W_2, R) = \frac{Cont\_frec(W_1, W_2)}{Cont\_frec(W_2, R)} \tag{2.10}$$

Se determina la probabilidad de aparición de la relación entre dos palabras W_1, W_2, con respecto a la probabilidad de aparición de esa misma relación R con respecto a una de las dos palabras en todo el corpus.

Relaciones palabra-clase (Resnik (1993)):

Se cuantifica la contribución de una clase semántica utilizando todos los conceptos que comparte esa clase.

$$A(W_1, C_2, R) = \frac{P(C_2|W_1, R) \log \frac{P(C_2|W_1,R)}{P(C_2)}}{\sum_{C_2} P(C_2|W_1, R) \log \frac{P(C_2|W_1, R)}{P(C_2)}} \tag{2.11}$$

Donde:

$$P(C_2|W_1, R) = \frac{Cont\_frec(W_1, C_2, R)}{Cont\_frec(W_1, R)} \tag{2.12}$$

$$Cont\_frec(W_1, C_2, R) = \sum_{W_2 \in C_2} \frac{Cont\_frec(W_1, W_2, R)}{Cont\_frec(W_2)} \tag{2.13}$$

Otros algoritmos utilizados para adquirir preferencias de selección son los propuestos por: (Agirre y Martinez (2001)) donde se utilizan relaciones clase-clase, como por ejemplo, "ingerir comida" es una relación clase-clase para "comer pollo" o también son utilizadas las redes bayesianas propuestas por (Ciaramita y Johnson (2000)).

Universitat d'Alacant
Universidad de Alicante

2.5.4.2 Usando preferencias de selección para WSD.

Una vez obtenidas las preferencias de selección éstas pueden ser integradas en un algoritmo de WSD de la siguiente forma.

1. Aprendizaje de un conjunto de preferencias de selección para una determinada relación sintáctica R
2. Dado un par de palabras $W_1 - W_2$ conectadas mediante una relación R
3. Encontrar todas las preferencias de selección $W_1 - C$ palabra-clase o $C_1 - C_2$ claseclase que se puedan aplicar
4. Seleccionar el sentido de W_1 y W_2 basados en la clase semántica elegida

Por ejemplo, si tratamos de desambiguar la palabra "café" en "beber café", los posibles sentidos de café son: 1. bebida, 2. árbol y 3. color. Utilizando la preferencia de selección "beber bebida" se seleccionaría el sentido café#1.

2.5.5 Heurísticas para Word Sense Disambiguation

Una forma sencilla de establecer el sentido correcto de las palabras en un texto es utilizar heurísticas basadas en propiedades lingüísticas aprendidas a través de textos. Una de las heurísticas más utilizadas como baseline es la denominada "sentido más frecuente". Además de esta heurística, existen otras dos comúnmente utilizadas cuya base es la suposición de que una palabra siempre tiene el mismo sentido en: todas sus ocurrencias en un mismo discurso ("un sentido por discurso") o en la misma colocación ("un sentido por colocación") o en el mismo dominio.

2.5.5.1 Sentido más frecuente.

Entre todos los posibles sentidos que puede tener una palabra, generalmente existe uno que ocurre más a menudo que los otros sentidos. Por lo tanto, un sistema muy simple de desambiguación sería aquel que asignara a cada palabra su sentido más frecuente.

Este método se utiliza a menudo como baseline para WSD, y de acuerdo a (Gale et al. (1992b)) "los sistemas deberían llegar como mínimo a este baseline".

En el gráfico de la Figura 2.3 se muestra la distribución de sentidos en el corpus Semcor. Los sentidos de cada categoría han sido obtenidos a partir de la distribución proporcionada por WordNet.

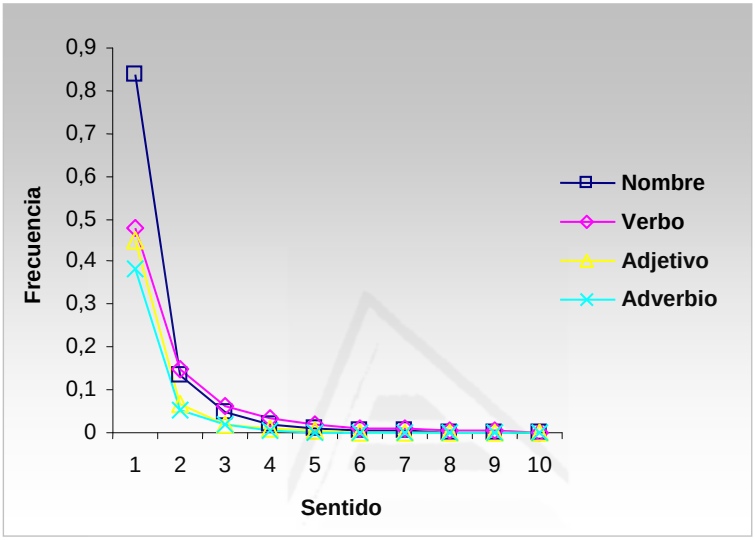


Figura 2.3. Distribución de sentidos en Semcor

Universitat d'Alacant
Universidad de Alicante

Como se puede apreciar, el sentido más frecuente para todas las categorías léxicas es el número 1.

Aunque conceptualmente es muy sencillo, y casi trivial de implementar, hay un inconveniente asociado a este método: no siempre disponemos de la distribución de las ocurrencias de los sentidos en todos los lenguajes, ya que, no existen suficientes textos disponibles para extraer esa distribución. Además, un cambio en el dominio generalmente altera la distribución de los sentidos, disminuyendo así los resultados obtenidos por esta heurística (Martínez y Agirre (2000)).

Para solventar el problema de la ausencia de textos que permitan obtener la distribución de los sentidos, existe una alternativa.

El método de (McCarthy et al. (2004)) propone la forma de utilizar una medida de similitud entre distintos sentidos de una palabra y palabras similares para determinar el sentido predominante en un dominio determinado.

El algoritmo utilizado por este método se compone de tres pasos:

1. Dada una palabra w encontrar las top k palabras similares. $N_w = \{n_1, n_2, ..., n_k\}$ con sus respectivos valores de similitud $\{dvs(w, n_1), dvs(w, n_2), ..., dvs(w, n_k)\}$
2. Para cada sentido ws_i de w, identificar la similitud con las palabras n_j usando el sentido de n_j que maximice el valor de similitud.
3. Ordenar los sentidos ws_i basándose en el valor de similitud total.

$$Similitud(ws_i) = \sum_{n_j \in N_w} dvs(w, n_j) \frac{wnvs(ws_i, n_j)}{\sum_{ws'_i \in sentidos(w)} wnvs(ws'_i, n_j)'} \tag{2.14}$$

Donde $wnvs(ws_i, n_j) = \max_{ns_x \in sentidos(n_j)} (wnvs(ws_i, ns_x))$

Por ejemplo, supongamos que queremos determinar el sentido de la palabra *"pipe"* en un texto determinado. Los posibles sentidos de pipe son:

pipe#1: tobacco pipe.
pipe#2: tube of metal or plastic.

Las palabras similares detectadas en el texto son las siguientes:
N={tube, cable, wire, tank, hole, cylinder, fitting, ...}

Para cada palabra N se calcula el valor de similitud con el sentido pipe#i (escogiendo el valor de similitud que maximiza el par).

pipe#1 - tube#3 = 0,3

Universidad de Alicante

pipe#2 - tube#1 = 0,6

Se establece el valor de similitud total de cada sentido de pipe#i:

similitud(pipe#1) = 0,25
similitud(pipe#2) = 0,73

Este método fue presentado en la tarea de *"English all-words"* de SENSEVAL-2 obteniendo un 64 % de precisión sobre los nombres.

2.5.5.2 Un sentido por discurso.

Esta heurística fue introducida por (Gale et al. (1992a)), donde se establece que una palabra tiende a preservar su sentido a través de todas sus ocurrencias en un discurso determinado. Esta medida permite establecer el sentido de una misma palabra identificándolo una única vez.

Esta heurística funciona bien con palabras que tienen sentidos bien diferenciados. En el caso en que tengamos palabras con sentidos con una diferencia muy sutil, este método obtiene peores resultados. En el estudio realizado por (Krovetz (1998)) se demostró que palabras polisémicas con sentidos muy similares, pueden tener más de un sentido por discurso. En concreto, utilizaron el corpus de Semcor, probando que el 70 % de las palabras en este corpus tenía un sentido por discurso.

2.5.5.3 Un sentido por colocación.

Esta heurística tiene una hipótesis similar a la heurística de un sentido por discurso, pero aplicada en un ámbito diferente. Fue presentada por (Yarowsky (1993)), y supone que una palabra tiende a tener el mismo sentido cuando se utiliza en la misma colocación. Es decir, las palabras cercanas dan pistas acerca del sentido de una palabra. Además, se ha determinado que este efec-

to es mayor para colocaciones adyacentes y empieza a decrecer cuando la distancia entre palabras aumenta.

Por ejemplo, la palabra *"plant"* en la colocación *"industrial plant"* mantiene su sentido en todas las ocurrencias, independientemente del contexto en el que aparezca esta colocación.

Se desarrollaron distintos experimentos con palabras con sentidos bien diferenciados y con sentidos muy próximos entre sí. Al igual que en el caso anterior, los resultados empeoran cuando se consideran palabras con sentidos con diferencias sutiles (Martínez y Agirre (2000)).

2.6 Métodos no supervisados basados en corpus

El desarrollo de métodos que tratan de resolver el problema de la ambigüedad léxica ha supuesto la aparición de diferentes algoritmos que utilizan una serie de recursos diferentes. Podemos encontrar desde sistemas que utilizan técnicas de enriquecimiento de conocimiento utilizando diccionarios, tesauros o jerarquías de conceptos (los llamados basados en conocimiento), hasta sistemas que utilizan la información de textos anotados semánticamente (los llamados sistemas supervisados basados en corpus). El único inconveniente de estos sistemas es que es necesario la creación de textos, diccionarios u otras fuentes de información, de forma manual. Esto supone un gran costo en su obtención y mantenimiento, además de generar dificultades cuando se tratan de anotar textos muy extensos, de un nuevo dominio o de un lenguaje diferente.

Para evitar esta dependencia se han desarrollado dos aproximaciones diferentes. La primera de ellas trata de establecer distinciones entre sentidos basándose en su distribución, determinando por tanto que, palabras que aparecen en contextos similares deben tener sentidos similares (Harris (1968), Miller y Charles (1991)). La segunda aproximación está basada en equivalencias de traducción en corpus paralelos, los cuales identifican traducciones de una palabra en un lenguaje determinado cuya traducción depende del sentido de esa palabra en el lenguaje origen. Estas traducciones

dependientes del sentido de una palabra pueden ser utilizadas como una recopilación de sentidos para esa palabra en el lenguaje origen.

Una de las claves de los métodos basados en distribución, es que no utilizan ningún recopilatorio de sentidos, únicamente clasifican palabras basándose en sus contextos observados en los corpus. Esta es una alternativa a los métodos que dependen de la anotación de corpus y que están restringidos a aquellas palabras que un experto ha clasificado para sus distintos sentidos. En todo caso, a pesar de que exista un repertorio de sentidos, su utilidad depende de las aplicaciones sobre las que se aplique.

Las aproximaciones distribucionales no asignan sentidos a las palabras, pero sí permiten discriminar entre los sentidos de una palabra identificando clusters en contextos similares, donde cada cluster muestra que una palabra se está utilizando con un sentido determinado. Estos métodos presentan una aproximación diferente a la tarea tradicional de WSD, la cual clasifica palabras con respecto a un repertorio de sentidos existente.

Los métodos basados en equivalencias de traducción se basan en el hecho de que los sentidos diferentes de una palabra en un lenguaje origen se pueden traducir en palabras diferentes en el lenguaje destino. Estas aproximaciones tienen dos propiedades. Primero, automáticamente derivan un repertorio de sentidos que hace distinciones relevantes para los problemas de traducción automática. Segundo, un corpus etiquetado basado en estas distinciones puede ser creado automáticamente y utilizado como corpus de entrenamiento para los métodos tradicionales de aprendizaje supervisado.

Una de las ventajas de utilizar métodos no supervisados basados en corpus, es que no se basan en ningún diccionario, repositorio de sentidos, tesauro, etc. De forma que no están restringidos a la interpretación de sentidos que el autor del diccionario haya impuesto. Ya que, es muy habitual que diferentes diccionarios aporten una distinción de sentidos más fina o más compacta, según la finalidad para la que estén creados. Al evitar hacer uso de estos recursos, se garantiza la adaptabilidad de estos sistemas a diferentes campos o ámbitos. Otra ventaja no menos importante,

es que estos métodos son independientes del lenguaje. Es decir, son fácilmente adaptables a cualquier idioma que disponga de un corpus sobre el que obtener información.

2.6.1 Métodos distribucionales

Este tipo de métodos identifican las palabras que suelen aparecer en contextos similares, sin necesidad de utilizar un repositorio de sentidos. En (Schütze (1998)) por ejemplo, se realiza el proceso de desambiguación en dos pasos. El primer paso, es construir clusters que comparten características similares. El segundo paso, es etiquetar cada cluster con una definición que establezca el sentido de la palabra dentro de ese contexto. Esta es una visión completamente diferente del concepto general de WSD, donde los sentidos se suponen conocidos antes de comenzar el proceso de desambiguación.

Esta nueva visión de "discriminación y etiquetación" corresponde a la forma ideal de obtener la definición de una palabra (lexicografía). Un lexicógrafo, seleccionaría diferentes contextos de una palabra determinada, a partir de un corpus extenso y representativo para el usuario final. Por ejemplo, si hablamos de un diccionario para niños, el corpus consistiría en textos escritos para niños. Y si hablamos de un diccionario sobre un dominio específico el corpus deberían ser textos de esa especialidad en particular. De esta forma el lexicógrafo, dividiría los contextos en los que aparece la palabra a estudiar en diferentes clusters, discriminando los diferentes sentidos que puede adoptar esa palabra, sin tener ninguna idea preconcebida de cuántos sentidos puede adoptar.

El resultado de la discriminación es un número determinado de clusters que establecen los diferentes sentidos de la palabra, obtenidos éstos a partir del corpus de entrada. A partir de aquí, se debe estudiar cada cluster y obtener una definición que actúe como una etiqueta o un sentido específico para la palabra. Esta última parte, la de asignar una definición concreta a la palabra en cada cluster es la más problemática, dado que en muchas ocasiones es difícil establecer una definición a partir de los contextos. Una posible solución sería identificar el conjunto de palabras que

aparecen en un cluster y utilizarlas como una aproximación al sentido de la palabra. Por ejemplo, si tenemos la palabra "línea" y un cluster con: "teléfono", "llamada", "ocupada", "móvil". En este caso, estas palabras son indicativas del sentido asociado a este cluster.

De esta forma, si los métodos no supervisados basados en corpus son desarrollados eficientemente, el resultado podría llegar a ser un proceso independiente del lenguaje que resuelve el problema de la ambigüedad sin tener que recurrir a un repositorio de sentidos.

Existen dos aproximaciones distintas para los métodos distribucionales:

Discriminación basada en tipos. Estos métodos identifican conjuntos (o clusters) de palabras que pueden estar relacionadas entre sí debido a su aparición en contextos similares. Normalmente se basan en medidas de similitud entre vectores de co-ocurrencia.

Discriminación basada en tokens. Estos métodos agrupan todos los contextos donde una palabra determinada aparece, basándose en la similitud de estos contextos.

2.6.1.1 Discriminación basada en tipos.

En el caso de los métodos de discriminación basados en tipos, es necesario disponer de corpus extensos para poder extraer la similitud entre diferentes contextos donde aparece la palabra a desambiguar. En estos métodos la representación más utilizada se basa normalmente en la contabilización de co-ocurrencias o en medidas de asociación entre palabras. Usando esta información es posible identificar otras palabras que aparecen en contextos similares y por tanto pueden tener sentidos similares. De esta forma, se pueden extraer los distintos sentidos que puede adoptar una palabra polisémica.

Por ejemplo, si seleccionamos la palabra "línea" que puede tener varios sentidos (línea telefónica, trazo, premio en el juego del bingo, etc), y ésta aparece en dos contextos distintos: **contexto1** (dibujo, trazo, color, coordenada) y **contexto2** (auricular,

teléfono, comunicar, llamada). Podemos establecer a partir de las palabras extraídas del contexto, que en el primer caso "línea" hace referencia a un trazo en un dibujo, y en el segundo caso, hace referencia a una línea telefónica.

Como ya se ha mencionado anteriormente, los métodos distribucionales basados en tipos necesitan de corpus bastante extensos. Es por ello, que la representación del espacio contextual se realizará en matrices de NxN dimensiones, donde N, es el número de palabras en el corpus. Cada celda de esta matriz contiene el número de veces que las palabras representadas en cada columna y fila co-ocurren dentro de una ventana de un tamaño especificado. Cuando no importa el orden en el que aparecen las palabras la frecuencia será la misma, pero si hablamos de bigramas, donde el orden sí importa, el valor de las celdas será distinto. Por tanto, si el orden no importa, se tendrá una matriz cuadrada y simétrica. Sin embargo, si tenemos en cuenta el orden de aparición de las palabras, tendremos una matriz rectangular y no simétrica.

Para estas matrices de co-ocurrencia, las celdas pueden almacenar el número de veces que dos palabras co-ocurren, o también pueden tomar valores más complejos. Por ejemplo, las celdas de una matriz de co-ocurrencia pueden contener el valor de diferentes medidas de asociación como: log-likelihood ratio (Rayson y Garside (2000)) o Información Mutua (Church y Hanks (1990)). Estas medidas indican el grado en que dos palabras co-ocurren con respecto a las demás palabras del corpus.

En el caso de la medida del log-likelihood ratio partimos de una tabla 2×2 como sigue a continuación 2.7.

	Corpus1	Corpus2	Total
Frecuencia de la palabra	a	b	$a+b$
Frecuencia de otras palabras	$c-a$	$d-b$	$c+d-a-b$
Total	c	d	$c+d$

Tabla 2.7. Tabla $2x2$ para log-likelihood ratio

En la Tabla 2.7 se extraen las frecuencias relativas de una palabra entre dos corpus. Se denota por c al número de palabras

total del corpus1 y por d al número de palabras total del corpus2 (N en total). Los valores de a y b son denotados como valores observados (O). Por último, queda por definir los valores esperados (E), según la Fórmula 2.15.

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i} \tag{2.15}$$

Para la Tabla 2.7 $N_1 = c$ y $N_2 = d$. Por lo tanto, para la palabra que estamos tratando:

$$E_1 = \frac{c * (a+b)}{(c+d)} \qquad y \qquad E_2 = \frac{d * (a+b)}{(c+d)} \tag{2.16}$$

Los cálculos para obtener los valores esperados tienen en cuenta el tamaño de los dos corpus. Por tanto, no es necesario normalizar los valores, pudiendo aplicar a continuación la medida del log-likelihood según la Fórmula 2.17 .

$$-2 \ln \lambda = 2 \sum_i O_i \ln \left(\frac{O_i}{E_i} \right) \tag{2.17}$$

La Fórmula 2.17 es equivalente a calcular el log-likelihood ratio G^2 como sigue:

$$G^2 = 2 * \left(a * \ln \left(\frac{a}{E_1} \right) \right) + \left(b * \ln \left(\frac{b}{E_2} \right) \right) \tag{2.18}$$

Si los valores esperados y los observados son comparables, el valor de G^2 estará próximo a 0, lo que significa que la palabra ha aparecido junto a otra por casualidad, y no están relacionadas entre sí. Si se obtiene un valor mayor que 0, significa que los valores observados difieren en gran medida de los valores esperados, por lo que las palabras estarán fuertemente relacionadas entre sí.

Una vez decidido el tipo de medida a utilizar para establecer la co-ocurrencia entre distintas palabras y construida la matriz de co-ocurrencia, cada palabra será representada como un vector de N-dimensiones. A partir de cada vector obtenido, se puede medir la similitud contextual entre dos palabras obteniendo el coseno entre los vectores. Para el cálculo del coseno entre dos vectores se utiliza la Fórmula 2.19.

$$\cos(\overrightarrow{x}, \overrightarrow{y}) = \frac{\overrightarrow{x} \cdot \overrightarrow{y}}{|\overrightarrow{x}| \times |\overrightarrow{y}|} \tag{2.19}$$

Continuando con la definición de métodos distribucionales basados en tipos, encontramos distintos algoritmos que pueden ser aplicados. En esta sección vamos a tratar dos de estos algoritmos: Análisis de la Semántica Latente (LSA) (Deerwester et al. (1990)) y Clustering by Committee (CBC) (Pantel y Lin (2002)).

Mediante el algoritmo de LSA se representa un corpus en un espacio multidimensional, usando vectores. Cada vector representará el contexto en el cual aparece una palabra. En el caso de LSA no se hacen distinciones entre los distintos sentidos de una palabra polisémica, es decir, se formará un único vector para cada palabra, aunque ésta tenga varios sentidos diferentes. Usando la información del contexto, se podrá determinar, por ejemplo, que palabras como: coche, automóvil, auto... están relacionadas semánticamente.

Cuando hablamos de LSA, debemos pensar en la representación del conocimiento como matrices de [palabras-contextos]. Para medir el grado de similitud de una palabra con respecto a otras palabras del contexto, se utiliza la medida del coseno entre vectores. Además de poder comparar palabras y contextos, también se puede medir el grado de similitud entre oración-oración, contexto-contexto... simplemente calculando el vector resultado de la unión de cada uno de los vectores que conforman las palabras de la oración, del contexto, etc.

Mediante el algoritmo de CBC se pueden detectar clusters de palabras relacionadas con los distintos sentidos de una palabra polisémica. Por ejemplo, para la palabra “muñeca” el algoritmo

de CBC podría identificar dos clusters, uno asociado con el sentido de juguete, con palabras como juego, entretenimiento, niños, cochecito, etc, y otro cluster asociado con el sentido de parte del cuerpo humano, con palabras como brazo, extremidad, articulación, etc. Por lo tanto, con el algoritmo de CBC se pueden detectar palabras sinónimas asociadas a los diferentes sentidos de una palabra.

Ambos algoritmos, tanto LSA como CBC, utilizan representaciones multidimensionales de co-ocurrencia de palabras.

2.6.1.2 Discriminación basada en tokens.

El objetivo de este tipo de métodos es agrupar los contextos en los que una palabra aparece bajo el mismo sentido.

A continuación se van a describir métodos que utilizan características de primer y segundo orden. Las características de primer orden ocurren directamente en un contexto que está siendo clasificado, mientras que las características de segundo orden son aquellas que ocurren junto con una de primer orden, pero no ocurren en el contexto que está siendo clasificado.

En primer lugar, es necesario establecer cómo se van a representar los contextos que van a ser clasificados en clusters. Al igual que para los sistemas supervisados, los contextos contienen la palabra a desambiguar con la excepción de que ésta no tiene asignado ningún sentido. La premisa de los métodos basados en tokens es que si una palabra aparece en contextos similares ésta ha de tener el mismo sentido.

Uno de los primeros métodos que utilizó discriminación basada en tokens fue una adaptación del algoritmo de LSA usando características de segundo orden (Schütze (1998)). En este caso, la representación de la matriz de co-ocurrencia en lugar de utilizar palabras utiliza contextos completos usando co-ocurrencias de segundo orden de características léxicas. Una palabra tiene una co-ocurrencia de segundo orden con otra, cuando ambas no aparecen juntas pero ambas sí aparecen junto a otra palabra frecuentemente. Por ejemplo, en “policía de tráfico” y “accidente de tráfico”, la palabra “policía” es una co-ocurrencia de segundo

orden de "accidente", porque ambas co-ocurren en primer orden con "tráfico".

Otro método que utiliza esta aproximación es el de (Pedersen y Bruce (1997)). En este caso, utilizan un conjunto reducido de características de primer orden para crear matrices que muestran la similitud entre contextos. Estas características se extraen a partir de las palabras que se encuentran alrededor de la palabra a desambiguar e incluyen etiquetas sintácticas y palabras co-ocurrentes.

El problema de este tipo de métodos es la forma de evaluación de los resultados. Debido a que la discriminación no parte de un conjunto preestablecido de sentidos, no se puede evaluar la forma en que los nuevos sentidos son descubiertos, sobretodo si se está trabajando en un dominio específico.

2.7 Métodos supervisados basados en corpus

Los métodos supervisados realizan la desambiguación de forma automática a partir de modelos o reglas obtenidas a partir de textos anotados previamente. Cuando hablamos de textos anotados, nos referimos a textos cuyo contenido ha sido etiquetado de forma manual. En este caso, la etiquetación se corresponde tanto a la parte de semántica como a la parte sintáctica. Ya que, como se comentó anteriormente se debe identificar el tipo de categoría sintáctica de una palabra, para poder establecer su sentido semántico.

Universidad de Alicante

En líneas generales los pasos a seguir por un método supervisado son los siguientes:

1. Seleccionar un conjunto de ejemplos que muestren las distintas clasificaciones de cada elemento.
2. Identificar patrones asociados a cada elemento.
3. Generalizar los patrones en reglas.
4. Aplicar las reglas para clasificar nuevos elementos.

Dentro de este tipo de métodos cabe destacar las técnicas basadas en Aprendizaje Automático (Machine learning systems) (Mit-

chell (1997b), Mitchell (1997a)). Estas técnicas han sido ampliamente utilizadas en tareas de PLN obteniendo un éxito considerable.

Los problemas iniciales de PLN sobre los que fueron aplicados este tipo de métodos estadísticos y de aprendizaje automático, fueron aquellos vinculados a la resolución de la ambigüedad léxica. En este tipo de tareas, se debe seleccionar de entre un conjunto de alternativas, la interpretación correcta para una palabra en un contexto determinado. Podemos destacar tareas tales como: selección de palabras en reconocimiento de voz, traducción automática, desambiguación automática, resolución de co-referencias, etc. Este tipo de tareas se consideran adecuadas para un sistema de aprendizaje automático porque pueden ser vistas como problemas de clasificación, donde el sistema de aprendizaje trata de etiquetar (clasificar) una serie de elementos, utilizando una de entre varias categorías (clases). En este caso, la base de conocimiento del sistema está formada por ejemplos previamente etiquetados.

Actualmente, las técnicas de aprendizaje automático han sido aplicadas a otros problemas de PLN, los cuales, no se reducen a un simple problema de clasificación. Dentro de estas nuevas aplicaciones encontramos: etiquetación de secuencias (con entidades, categorías sintácticas, etc) y asignación de estructuras jerárquicas (árboles sintácticos, conceptos complejos en extracción de información, etc). En estos casos, se parte de un problema complejo que puede ser descompuesto en esquemas de decisión simples o se pueden generalizar los conjuntos de clasificación para trabajar directamente con representaciones y salidas complejas.

En relación a WSD, en los últimos diez años, la técnica de aprendizaje supervisado, a partir de ejemplos, ha sido una de las que mejores resultados ha obtenido. En este caso, los modelos estadísticos o de clasificación se obtienen a partir de corpus anotados semánticamente. Normalmente, los métodos supervisados han obtenido mejores resultados que los no supervisados. Esta afirmación queda demostrada a la vista de los resultados conseguidos en las últimas competiciones realizadas para la evaluación de métodos de análisis semántico (ACL (2001),ACL (2004)). Sin embargo, estos métodos tienen un grave problema, la necesidad

de disponer de corpus lo bastante extensos para poder entrenar los sistemas. A menudo, escasean los corpus anotados debido a su costoso proceso de anotación manual, es el conocido problema del cuello de botella de la adquisición de conocimiento. Esta restricción afecta en gran medida a los sistemas, ya que no tienen la materia prima necesaria para poder trabajar.

2.7.1 El proceso de clasificación en aprendizaje supervisado

El objetivo principal en el aprendizaje supervisado para la tarea de clasificación consiste en inducir a partir de un conjunto de entrenamiento C, una aproximación (o hipótesis) h de una función no conocida f que mapea a partir de un espacio de entrada E a un espacio de salida $S = 1, ..., K$.

El conjunto de entrenamiento contiene m ejemplos de entrenamiento, $C = (\overrightarrow{e}^1, y^1), ..., (\overrightarrow{e}^m, y^m)$, pares $(\overrightarrow{e}, y)$. Donde $\overrightarrow{e}$ pertenece a E y $y = f(\overrightarrow{e})$. El componente $\overrightarrow{e}$ de cada ejemplo es normalmente un vector $\overrightarrow{e} = (e_1, ..., e_n)$, cuyos componentes, llamados atributos (features) describen información relevante acerca del ejemplo. Los valores del espacio de salida S asociados con cada ejemplo de entrenamiento se llaman clases (categorías). Por lo tanto, cada ejemplo de entrenamiento está completamente descrito por un conjunto de pares atributo-valor y una etiqueta de clase.

Según la teoría del aprendizaje estadístico (Vapnik (1998)), la función f se considera como una función de distribución de probabilidad $P(X, Y)$ y los ejemplos de entrenamiento se consideran como una muestra de esa distribución. Además, X se identifica normalmente con $\Re^n$, y cada ejemplo $\overrightarrow{x}$ como un punto en $\Re^n$ con un valor real en cada dimensión. Estas son las dos posibles notaciones que podemos encontrar en este tipo de sistemas.

Dado un conjunto de entrenamiento C, un algoritmo de aprendizaje induce un clasificador denotado como h, el cual es utilizado como una hipótesis sobre la verdadera función f. A partir de aquí el algoritmo de aprendizaje puede seleccionar entre un conjunto de posibles funciones H, a las que se llama *espacio de*

hipótesis. Los algoritmos de aprendizaje se diferencian en base a dos rasgos: el tipo de espacio de hipótesis que manejan: funciones lineales, funciones radiales, etc. O el tipo de algoritmo de selección que utilizan para decidir cuál de las hipótesis es la mejor con respecto al corpus de entrenamiento: simplicidad, margen máximo, etc.

Dados nuevos vectores $\overrightarrow{x}$, h se utiliza para predecir los correspondientes valores y. En este caso, se clasifican los nuevos ejemplos, y el resultado se prevee que coincida con f en la mayoría de los casos, o de forma equivalente, que conlleve al menor número de errores. La forma de estimar el grado de error en aquellos ejemplos nunca vistos anteriormente se denomina *error de generalización*. Este tipo de errores no pueden ser minimizados por el algoritmo de aprendizaje, dado que la función f o la distribución $P(X, Y)$ son desconocidas. Por lo tanto, es necesario un principio de inducción. La forma más común de proceder es minimizar el denominado *error de entrenamiento*, es decir, el número de errores que encontramos en el conjunto de entrenamiento. Esta acción se conoce como la *minimización del riesgo empírico* y proporciona una buena estimación del error de generalización con los suficientes ejemplos de entrenamiento. Sin embargo, para dominios con pocos ejemplos de entrenamiento, podemos ajustar demasiado los datos de entrenamiento y generalizar erróneamente. El riesgo de ajuste se ve incrementado cuando tenemos datos atípicos y ruido.

2.7.1.1 Ejemplo: WSD con aprendizaje automático.

Supongamos que se quieren desambiguar las diferentes ocurrencias del verbo *"to know"* en diferentes contextos. En este caso, se considerarán los diferentes sentidos del verbo como las distintas clases del problema de clasificación (espacio de salida Y). Además, cada ocurrencia del verbo en un corpus previamente anotado semánticamente, será codificada como un ejemplo (x^i) para la tarea de entrenamiento. En la Tabla 2.8 el verbo *"to know"* tiene once sentidos diferentes según las definiciones de WordNet 1.6, con sus correspondientes dominios de WordNet Domains.

Synset	Dominio	Glosa
00401762 know#1	psychology	be cognizant or aware of a fact or a specific piece of information; possess knowledge or information about; "I know that the President lied to the people"; "I want to know who is winning the game!"; "I know it's time"
00402497 know#2	psychology	know how to do or perform something; "She knows how to knit"; "Does your husband know how to cook?"
00402210 know#3	psychology	be aware of the truth of something; have a belief or faith in something; regard as true beyond any doubt; "I know that I left the key on the table"; "Galileo knew that the earth moves around the sun"
00401559 know#4	factotum	be familiar or acquainted with a person or an object; "She doesn't know this composer"; "Do you know my sister?" "We know this movie"
00402992 know#5	psychology	have firsthand knowledge of states, situations, emotions, or sensations; "I know the feeling!" "have you ever known hunger?"
00400501 know#6	factotum	discern; "His greed knew no limits"
00402658 know#7	factotum	have fixed in the mind; "I know Latin"; "This student knows her irregular verbs"; "Do you know the poem well anough to recite it?"
00977560 know#8	sexuality	have sexual intercourse with; "This student sleeps with everyone in her dorm"; "Adam knew Eve" (know is archaic); "Were you ever intimate with this man?"
00411402 know#9	psychology	know the nature or character of; "we all knew her as a big show-off"
00411252 know#10	factotum	be able to distinguish. recognize as being different; "The child knows right from wrong"
00411122 know#11	factotum	perceive as familiar; "I know this voice!"

Tabla 2.8. Sentidos del verbo "to know" en WordNet 1.6

Generalmente, las definiciones de los sentidos de una palabra, tienen asociados ejemplos con información relevante del contexto donde suele utilizarse. Esta información puede usarse para extraer características (*"features"*), como por ejemplo bigramas, trigramas, relaciones sintácticas, etc. Estas características son utilizadas para codificar los ejemplos de entrenamiento mediante vectores de n dimensiones, donde n es el número de características

utilizado. En el caso de los algoritmos de aprendizaje automático, es imprescindible obtener la información del contexto que rodea a la palabra ambigua para poder construir los vectores de características. Normalmente, es necesario realizar un pre-proceso para poder construir estos vectores. Es preciso por una parte, obtener las palabras con contenido semántico que rodean a la palabra ambigua, para ello, se establecen ventanas contextuales de diferente tamaño, también se utilizan analizadores sintácticos para estudiar los patrones de relaciones sintácticas, se detectan las palabras compuestas, etc. Este pre-proceso es necesario para una correcta definición de las características y determinará el buen funcionamiento del algoritmo de aprendizaje automático.

Los conjuntos de características más utilizados en aprendizaje automático se pueden clasificar en tres grupos:

Características locales Las características locales engloban: n-gramas de etiquetas sintácticas, lemas, palabras junto con su posición respecto a la palabra a desambiguar. En alguna ocasión las características locales incluyen sacos de palabras o lemas situados en el entorno de la palabra ambigua. Mediante estas características se puede capturar el conocimiento sobre colocaciones, relaciones sintácticas, etc.

Características generales Mediante las características generales se pueden representar contextos mucho más generales. La representación de estas características se realiza mediante sacos de palabras (ventana amplia de palabras, oraciones, párrafos, documentos...). Usando este tipo de características se puede capturar el dominio semántico de un fragmento de texto o de un documento.

Dependencias sintácticas Al nivel de una oración, las dependencias sintácticas se pueden utilizar para modelar relaciones entre diferentes argumentos.

Además de los vectores de características también se suelen utilizar listas de decisión. En este caso, el algoritmo de aprendizaje se basa en una serie de reglas del tipo:

$if\ (caracteristica\ =\ valor)\ then\ clase$

En el caso de algoritmos basados en listas de decisión, cada vez que se trata de clasificar un nuevo ejemplo **x**, se van ejecutando por orden el listado de reglas hasta que se encuentre una que se pueda aplicar sobre el nuevo ejemplo.

Suponiendo que se han obtenido una serie de reglas de decisión a partir de varios ejemplos de entrenamiento, se podría ejecutar un algoritmo de decisión sobre el siguiente ejemplo: *"There is nothing in the whole range of human experience more widely known and universally felt than spirit"*. La Tabla 2.9 muestra las reglas aplicadas junto con una probabilidad de certidumbre para cada regla.

Característica	Valor	Sentido	Probabilidad
± ventana de 3 palabras	"widely"	4	2,99
bigrama	"widely known"	4	2,99
bigrama	"known and"	4	1,09
ventana	"whole"	1	0,91
ventana	"widely"	4	0,69
ventana	"known"	4	0,43

Tabla 2.9. Clasificación según listas de decisión de la palabra *"know"*

En este ejemplo, se puede observar que la lista de decisión únicamente establece valores positivos para los sentidos 1 y 4 de *"know"*. Usando esta información, se podría proponer como sentido correcto de *"know"* el 4, debido a que la mayoría de reglas apuntan a este sentido.

Para finalizar, recordar que cuando se habla de "aprendizaje supervisado" se parte de un corpus de entrenamiento previamente anotado en base a una serie de clases semánticas. Mientras que cuando se habla de "aprendizaje no supervisado" no existe anotación previa y el objetivo final es, a partir de similitudes semánticas obtener una serie de clusters para poder ser interpretados como clases semánticas.

Universidad de Alicante

2.7.2 Clasificación de métodos de aprendizaje supervisado

A continuación se van a describir algunos de los distintos métodos de aprendizaje supervisado utilizados en WSD. Estos métodos son clasificados atendiendo a la forma que tienen de adquirir los modelos de clasificación.

2.7.2.1 Métodos probabilísticos.

Los métodos estadísticos normalmente estiman un conjunto de parámetros que determinan la probabilidad condicional de las categorías y los contextos (descritos mediante características). Estos parámetros se utilizan para asignar a cada nuevo ejemplo una categoría que maximice la probabilidad condicional a partir de las características observadas anteriormente.

El clasificador más simple que existe es el denominado Naïve Bayes Classifier (NBC) (Duda et al. (2001)). En este modelo, hay un nodo que representa la variable de clase C y un nodo para cada atributo x_i del ejemplo (ver Figura 2.4). Se parte de la hipótesis de que los valores de los atributos se generan independientemente a partir de la clase C de acuerdo con las distribuciones individuales $P(x_i|C)$. Para predecir la clase de un ejemplo, se elige la que maximiza la probabilidad de haber generado el ejemplo observado. Para ello, se utiliza una fórmula derivada a partir del teorema de Bayes. Este algoritmo ha sido usado en distintas tareas de PLN para resolver diversos problemas (categorización de documentos (Lewis y Ringuette (1994)), corrección ortográfica (Golding (1995)), resolución de la ambigüedad semántica (Leacock et al. (1998), Escudero et al. (2000)) ...) y a pesar de su extrema simplicidad, ha obtenido resultados notables. Además, utilizando el NBC se puede combinar información estadística de distintas fuentes, siempre que sean independientes.

La fórmula general para obtener la clasificación según el clasificador bayesiano es la siguiente:

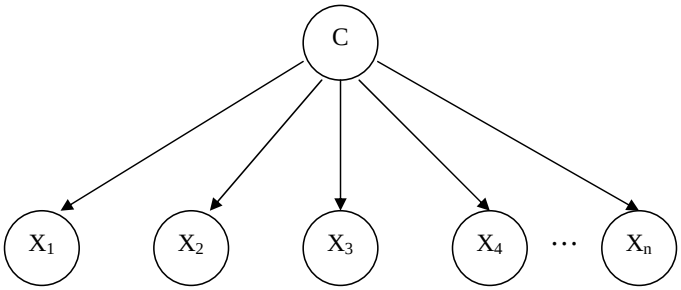


Figura 2.4. Modelo clasificador bayesiano (naïve)

$$P(C|X_1, X_2, X_3, ..., X_n) = \frac{P(X_1, X_2, X_3, ..., X_n|C) \times P(C)}{P(X_1, X_2, X_3, ..., X_n)} \tag{2.20}$$

Dado que los atributos son independientes con respecto a la clase C se cumple que:

$$P(X_1, X_2, X_3, ..., X_n|C) = \prod_i P(X_i|C) \tag{2.21}$$

De esta forma, el clasificador bayesiano naïve obtiene el siguiente resultado:

$$valor = \arg \max_{valor \in C} P(X_1|C) \times ... \times P(X_n|C) \times P(C) \tag{2.22}$$

Para el caso de WSD el sentido correcto para una palabra cualquiera C sería aquel que hiciera máximo el resultado de la Ecuación 2.22.

Por ejemplo, supongamos que tenemos 2000 instancias de la palabra *"bank"*: 1500 para bank#1 (financial) y 500 para bank#2 (river). En este caso, las probabilidades para cada sentido serían:

$P(S = 1) = 1500/2000 = 0{,}75$
$P(S = 2) = 500/2000 = 0{,}25$

Universitat d'Alacant
Universidad de Alicante

Dada la palabra *"credit"* ésta aparece 200 veces con bank#1 y 4 veces con bank#2.

$P(X_1 = credit) = 204/2000 = 0{,}102$
$P(X_1 = credit|C = 1) = 200/1500 = 0{,}133$
$P(X_1 = credit|C = 2) = 4/500 = 0{,}08$

Dado un texto que contiene la palabra *"credit"*:

$P(C = 1|X_1 = credit) = (0{,}133 \times 0{,}75)/0{,}102 = 0{,}978$
$P(C = 2|X_1 = credit) = (0{,}08 \times 0{,}25)/0{,}102 = 0{,}20$

Por tanto, se deduciría que el sentido correcto para *"bank"* es el número 1.

La efectividad del clasificador bayesiano *"naive"* ha sido probada en diferentes estudios (Mooney (1996), Pedersen (1997)) que demuestran que este clasificador es tan bueno como cualquier otro método.

2.7.2.2 Métodos basados en reglas de discriminación.

Este tipo de métodos utilizan las llamadas listas de decisión (Rivest (1987)) o árboles de decisión (Quillian (1986), Quillian (1993)) donde se utilizan reglas asociadas a cada uno de los diferentes sentidos de una palabra. En este caso, dado un ejemplo a clasificar, el sistema selecciona una o más reglas que son satisfechas por las características del ejemplo y asigna un sentido basándose en sus predicciones.

Concretamente, una lista de decisión es un conjunto ordenado de reglas de la forma (condición, clase, peso). Un ejemplo de este tipo de listas se encuentra en la Sección 2.7.1.1. Una lista de decisión con reglas a las que se le asignan pesos establece que las reglas con condiciones excepcionales se sitúan al principio de la lista con un peso elevado, las reglas con condiciones generales se sitúan al final con un peso bajo y la última condición de la lista es una condición por defecto que acepta el resto de casos no contemplados. Los pesos se establecen de acuerdo a una función que mide el grado de asociación entre la condición y una categoría

particular a partir de un corpus de entrenamiento. Para clasificar un nuevo ejemplo, cada regla de la lista se comprueba secuencialmente y la categoría de la primera regla que cumple la condición se asigna al nuevo ejemplo.

En (Yarowsky (1994b) se utilizan listas de decisión para resolver un tipo específico de ambigüedad: los acentos en español y francés. En un trabajo posterior se aplicaron listas de decisión para WSD (Yarowsky (1995)). En este estudio, cada condición de la lista se correspondía con una característica (*"feature"*), donde los valores eran los sentidos de las palabras y los pesos se calculaban de acuerdo a una fórmula que estimaba la probabilidad de un sentido con respecto a una determinada característica.

En (Martínez et al. (2002), Martínez (2004)) las listas de decisión se emplean en un sistema de WSD junto con una serie de nuevas características sintácticas (relaciones gramaticales instanciadas, relaciones gramaticales...) y semánticas (modelos de preferencias de selección). Este sistema se ha empleado para desambiguar textos en Euskera e Inglés.

En el caso de los árboles de decisión las reglas de clasificación se generan en forma de una estructura n-aria de ramas de un árbol. Cada rama de un árbol de decisión representa una regla que comprueba un conjunto de características (nodos internos) y hace una predicción de la clase del nodo terminal. Este tipo de estructuras no se han empleado frecuentemente en WSD. En (Mooney (1996)) se utilizó el algoritmo de (Quinlan (1993)) y se realizó un estudio comparativo con varios algoritmos de aprendizaje automático para WSD. Este estudio concluyó que los árboles de decisión no estaban entre los algoritmos que mejor resolvían el problema de WSD. A pesar de ello, en Senseval-1 (Yarowsky (2000a)) presentó un sistema modificado de listas de decisión con algunas ramas condicionales que obtuvo muy buenos resultados en la tarea *"English Lexical Sample"*.

2.7.2.3 Bootstrapping.

Como se ha comentado anteriormente, el problema de los métodos basados en aprendizaje automático es la escasez de cor-

pus anotados semánticamente. Para evitar este problema existe un método que requiere de un mínimo conjunto de elementos anotados (sistemas mínimamente supervisados), es el denominado método "de semilla" o bootstrapping (Abney (2002), Abney (2004)). La idea de este método es que a partir de un mínimo conjunto de ejemplos anotados se pueden realizar sucesivos aprendizajes que se alimentan incrementalmente con el conocimiento adquirido en el anterior. El término "semilla" proviene del inicio de tal proceso iterativo, que no necesita más que una mínima cantidad de conocimiento previo para comenzar el aprendizaje.

Existen diferentes aproximaciones del método de bootstrapping: co-training y self-training.

La idea básica para ambas aproximaciones es la siguiente:

Se parte de un conjunto EE de Ejemplos de Entrenamiento etiquetados y de un conjunto EN de Ejemplos No Etiquetados. Se dispone de C_i Clasificadores.

Paso 1. Crear un conjunto de ejemplos NE', eligiendo P ejemplos aleatorios de NE.

Paso 2. Bucle de I iteraciones:

- Entrenar los clasificadores C_i sobre el conjunto etiquetado EE y etiquetar el conjunto no etiquetado NE'.

- Seleccionar los M mejores ejemplos y añadirlos al conjunto EE, manteniendo la distribución de EE.

- Rellenar el conjunto NE' con ejemplos de NE, manteniendo NE' en un tamaño constante P.

Un ejemplo del método co-training lo encontramos en (Blum y Mitchell (1998)) donde se utilizan dos clasificadores. Y un ejemplo del método self-training lo encontramos en (Nigam y Ghani (2000)) donde se utiliza un único clasificador.

2.7.2.4 Métodos basados en redes neuronales.

Otro tipo de métodos utilizados para WSD son aquellos basados en redes neuronales, algoritmos genéticos, etc (Veronis y Ide (1990), Towell y Voorhees (1998)). Una Red Neuronal Artificial (RNA) es un modelo de procesamiento de información que está inspirado en un sistema nervioso biológico (Group (1986)).

Éste se compone de un gran número de elementos de procesamiento interconectados (neuronas) trabajando conjuntamente para resolver problemas específicos. Las RNA aprenden con ejemplos, es por tanto necesaria la utilización de un proceso de aprendizaje para su configuración. La principal ventaja de las RNA radica en la resolución de problemas demasiado complejos para tecnologías convencionales, problemas que no tienen un algoritmo de solución específico o que es muy difícil de encontrar. Entre estos problemas se encuentran el reconocimiento de patrones y pronósticos, clasificación de datos, optimización... Sin embargo, en el ámbito del procesamiento del lenguaje natural aún no han sido suficientemente explotadas (Valdivia et al. (2002)).

En (García (2006)) se propone un sistema de WSD basado en el modelo de red neural de Kohonen (Kohonen (1989)), en su variante de Aprendizaje por Cuantificación Vectorial (Learning Vector Quantification o LVQ). Como recursos lingüísticos para el aprendizaje de la red se utiliza el corpus de Semcor, que está etiquetado con los sentidos de WordNet y el conjunto de párrafos artificiales generado a partir de todas las relaciones de WordNet. Además, integra el modelo de espacio vectorial (con vectores obtenidos a partir de Semcor) con LVQ para definir las categorías de la red. Este sistema participó en la tarea English Lexical Sample de Senseval-2 obteniendo una precisión del 59 %.

Universitat d'Alacant
Universidad de Alicante

2.8 Métodos híbridos

Los métodos que se encuentran dentro de este grupo son aquellos que no pueden englobarse exactamente dentro de los grupos anteriores. Es decir, son aquellos que utilizan en el proceso de desambiguación tanto fuentes de conocimiento externas como corpus anotados o no anotados.

Un método que combina la utilización de diccionarios con corpus no anotados es el ideado por (Luk (1995)). Este método utiliza las definiciones de LDOCE para extraer las palabras que identifican cada sentido, construyendo así para cada sentido una lista de palabras representativas. Utilizando esta información y las oracio-

nes del Brown Corpus[2] (Francis y Kucera (1979)) no anotado, se expande el conjunto de palabras representativas obtenido a partir de LDOCE, de la siguiente forma: se extraen pares de palabras de cada oración del Brown Corpus y se determinan los conceptos co-ocurrentes mediante un algoritmo que obtiene una tabla de datos conceptuales co-ocurrentes. Esto permite producir un sistema que utiliza la información de recursos léxicos como un medio para reducir la gran cantidad de texto necesaria de los corpus de entrenamiento.

El proceso de desambiguación de una palabra polisémica W sobre un contexto C (la oración que contiene a la palabra), comienza dando valores a cada sentido S de W, según la fórmula:

$$score(S, C) = score(CS, C') - score(CS, GlobalCS) \qquad (2.23)$$

Donde CS es el conjunto de palabras representativas de LDOCE pertenecientes al sentido S, C' es el conjunto ampliado de palabras representativas y GlobalCS contiene las definiciones de cada concepto. A partir de estos valores y utilizando la Información Mutua entre las diferentes conjuntos de palabras representativas, se selecciona el sentido con mayor valor de Información Mutua.

Este sistema obtuvo un 77 % de precisión sobre las 12 palabras utilizadas en el trabajo de (Yarowsky (1992)).

Otros métodos destacables de este tipo son los publicados en: McRoy (1992), Dagan et al. (1991) y Dagan et al. (1994).

Existen también otros métodos que utilizan la combinación de tesauros y corpus no anotados, como es el caso del método ideado en (Yarowsky (1992)). Este método emplea la técnica de bootstrapping utilizando las palabras de las categorías del Roget's Thesaurus[3], considerándolas etiquetadas semánticamente y esta información se va aumentando utilizando un corpus no anotado.

Además podemos encontrar también métodos que combinan diferentes fuentes léxicas estructuradas con corpus: Lin (1997),

[2] http://nora.hd.uib.no/whatis.html
[3] http://www.gutenberg.org/etext/22

Agirre y Martínez (2000). Métodos que combinan WordNet y corpus: Resnik (1995b), Stetina et al. (1998), etc

En definitiva, el número de sistemas de WSD surgidos a partir de la combinación de diferentes fuentes de conocimiento es muy amplio y sería imposible citarlos exhaustivamente.

2.9 Otra clasificación de sistemas WSD

Dada la gran cantidad de métodos propuestos actualmente para WSD existe una clasificación más general que engloba únicamente dos tipos de sistemas: sistemas supervisados y sistemas no supervisados.

Esta clasificación es la utilizada en la competición SENSEVAL para la evaluación de los distintos sistemas de WSD presentados. Como ya se ha comentado en los puntos anteriores, cuando se habla de sistemas supervisados se hace referencia a aquellos sistemas que necesitan de corpus de entrenamiento anotados semánticamente. En cambio, los sistemas no supervisados son aquellos que no necesitan esa anotación para poder funcionar correctamente.

2.10 Aplicaciones actuales

Aunque no deja de ser importante en el Procesamiento del Lenguaje Natural, WSD se considera una tarea intermedia (Wilks y Stevenson (1998)), al igual que otras tareas como: part-of-speech tagging o análisis sintáctico. Decimos que es una tarea intermedia porque sus resultados únicamente proporcionan información lingüística y nada tienen que ver con lo que el usuario final demanda en última instancia. Otras tareas, las llamadas tareas finales, como la traducción automática, extracción de información y sistemas de diálogo, ofrecen unos resultados requeridos por el usuario, al que poco le importa el fondo lingüístico sobre el que este tipo de tareas se apoya. La mayoría de estas tareas finales requieren diferentes módulos que implementen una serie de tareas intermedias necesarias para el correcto funcionamiento de la aplicación.

Esencialmente, existen dos tareas finales que realmente obtienen beneficios por la utilización de un módulo de WSD. Estamos hablando de traducción automática y recuperación de información. A pesar de que en recuperación de información no están demostrados los beneficios de aplicar WSD sí se hace patente la necesidad de su utilización. Por ejemplo, cuando realizamos una búsqueda sobre el "águila imperial", queremos únicamente aquellos documentos que hablen sobre el ave rapaz, y no sobre peces de la especie raya o sobre un tipo de moneda española o mexicana. En 1992 y 1997 (Krovets y Croft (1992), Krovets (1997)), en unos estudios realizados con un corpus desambiguado manualmente, se comprobó que un sistema de WSD podría mejorar la recuperación de información en un 2 %. Otro estudio similar fue el realizado por Sanderson (Sanderson (1994)), en este caso, se efectuaron una serie experimentos similares a los realizados por Krovets. La única diferencia fue que en este caso la ambigüedad fue introducida artificialmente en los textos. En este caso, se demostró que el funcionamiento del sistema mejoraba para aquellas consultas que contenían menos de cinco palabras. En (Pekar et al. (2006)) se realiza un estudio para mejorar la traducción de palabras poco frecuentes a partir de una extensión de la medida de similitud de co-ocurrencia de palabras (Dagan et al. (1999)). Sin embargo, otros autores han corroborado que WSD contribuye de forma satisfactoria a mejorar los resultados de un sistema de RI. Esta afirmación viene corroborada por Schütze y Pedersen (Schütze y Pedersen (1995)), quienes demostraron que con la aplicación de WSD mejoraban el sistema de RI en un 14 %. Otros autores (Jing y Tzoukermann (1999)) también han demostrado que WSD mejora hasta en un 8,6 % los resultados en la recuperación de información. En este caso, utilizan un algoritmo de desambiguación que evalúa la similitud en el contexto local de la consulta, la similitud de la información en el corpus y las relaciones morfológicas entre palabras.

Además de poder ayudar a las tareas anteriormente mencionadas, WSD también puede ser muy útil para otras tareas. En el estudio realizado por Yarowsky (Yarowsky (1996)) se demuestra cómo WSD puede utilizarse para sistemas de síntesis de habla

o para encontrar la pronunciación correcta de homófonos, como por ejemplo, *"lead"* y *"live"*. También (Connine (1990)) demuestra que un sistema de WSD puede ayudar en el reconocimiento de voz a identificar el item léxico correcto para palabras con idénticas propiedades fonéticas, como "base" y *"bass"* o *"sealing"* y *"ceiling"*. Más allá de todas estas aplicaciones WSD también se puede utilizar en otras muchas tareas de procesamiento de textos (Yarowsky (1994b), Yarowsky (1994a)).

En la Tabla 2.10 se muestran algunas de las aplicaciones actuales que pueden obtener beneficios tras utilizar un buen sistema de WSD.

Aplicación	Ejemplo de uso de WSD
Traducción automática	Traducción de la palabra "bill" Inglés-Español Resultado: ¿Es un "pico" o una "cuenta"?
Recuperación de información	Encontrar todas las páginas web que hablen de "cricket" Resultado: ¿El deporte o el insecto?
Búsqueda de respuestas	¿Cuál es la opinión de George Miller sobre el control de armas? Resultado: ¿El psicólogo o el congresista?
Adquisición de conocimiento	Añadir a la base de conocimiento: Herb Bergson es el alcalde de Duluth Resultado: ¿Minnesota o Georgia?

Tabla 2.10. Utilización de WSD en aplicaciones de PLN

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Capítulo 3

Problemática en la evaluación de sistemas de WSD

En este capítulo se describe la problemática asociada a la evaluación de sistemas de desambiguación automática. Los problemas en esta tarea están centrados en el tipo de anotación utilizado para etiquetar los sentidos de cada palabra, los corpus utilizados, el criterio de selección del sentido correcto de una palabra y las medidas de evaluación utilizadas.

Universitat d'Alacant
Universidad de Alicante

3.1 Contexto del problema

En (Wilks y Stevenson (1998)) Yorick Wilks y Mark Stevenson, afirman que existen diferentes niveles y tipos de desambiguación, dependiendo de la proporción de palabras desambiguadas y el tipo de anotación utilizado. Además, para la tarea de desambiguación automática se utilizan distintos tipos de fuentes de información desde los Machine Readable Dictionaries (MRDs) como LDOCE y WordNet hasta los corpus anotados manualmente. Cada clase de sistema de WSD junto con los recursos que utiliza requiere un corpus de evaluación diferente. Por ejemplo, un corpus anotado en base a un diccionario particular no podrá ser utilizado por otro sistema que asigna los sentidos de otro diccionario distinto. Este

tipo de problemas derivan en una escasez de evaluación comparativa entre distintos sistemas.

Para tratar de unificar criterios de evaluación y establecer una comparativa entre distintos sistemas siguiendo los mismos estándares, se creó una competición llamada SENSEVAL (Evaluation Exercises for the Semantic Analysis of Text). Su primera edición fue en 1998 donde se evaluaron distintos sistemas en distintas lenguas: Inglés, Francés e Italiano. A partir de los resultados obtenidos y viendo el interés creado tras la competición, se han realizado nuevas ediciones (cada 3 años), para evaluar la evolución de los sistemas de desambiguación automática.

A continuación se describirán las dificultades para la evaluación y comparación de sistemas encontradas hasta la fecha y que han dado lugar a la organización de SENSEVAL.

3.1.1 Mejoras en los criterios de evaluación

Previamente a SENSEVAL la evaluación estándar de los sistemas de desambiguación automática era un simple mapeo exacto entre la anotación obtenida por el sistema y los sentidos correctos. De forma que la fórmula utilizada era la siguiente:

$$correctas\,\% = 100 \times \frac{\#\ sentidos\ anotados\ correctamente}{\#\ sentidos\ anotados\ total} \quad (3.1)$$

Sin embargo, esta forma de evaluación no es del todo adecuada teniendo en cuenta que un sistema puede devolver una probabilidad distinta para cada uno de los sentidos de una palabra polisémica. Por ejemplo, considerando el siguiente fragmento:

*"... bought an **interest** in Lydak Corp ..."*

Supongamos que la palabra ambigua *"interest"* es desambiguada por cuatro sistemas distintos, los cuales, asignan las siguientes probabilidades a los distintos sentidos de *"interest"* (ver Tabla 3.1).

Tal y como se aprecia en la Tabla 3.1, todos los sistemas seleccionan el sentido incorrecto de *"interest#1"* en lugar de *"interest#2"*. Sin embargo, el Sistema 1 es capaz de dar una probabilidad bastante elevada para el sentido correcto #2. A pesar de ello,

Sentido	Sistema 1	Sistema 2	Sistema 3	Sistema 4
#1 monetary (e.g. on a loan)	47	85	28	100
#2 stake or share ⇐ correcto	42	5	24	0
#3 bebefit/advantage/sake	6	5	24	0
#4 intellectual curiosity	5	5	24	0

Tabla 3.1. Distribución de probabilidades asignadas por diferentes sistemas

con la forma de evaluación dada en la Ecuación 3.1 el Sistema 1 se penaliza de igual forma que el resto de sistemas, aún teniendo en cuenta que ha sido capaz de predecir el sentido correcto con una probabilidad bastante alta.

Para evitar que los sistemas que obtienen distintas probabilidades para los sentidos de una palabra se penalicen de igual forma que los que sólo seleccionan un único sentido, se utiliza otro tipo de medida adaptable a todo tipo de sistemas: la entropía cruzada. En este caso, se evalúa la efectividad de las predicciones de un sistema con distintas probabilidades para los distintos sentidos. La fórmula de la entropía cruzada es la siguiente:

$$Entropia\ cruzada = -\frac{1}{N}\sum_{i=1}^{N}\log_2 Pr_S(sc_i|w_i, context_i) \quad (3.2)$$

Donde N es el número de instancias de test y Pr_s es la probabilidad asignada por el sistema S al sentido correcto sc_i de la palabra w_i en el contexto $context_i$. Mediante esta nueva forma de evaluación, los sistemas que asignen una probabilidad bastante elevada al sentido correcto, obtendrán mejores resultados que los demás.

La entropía cruzada como medida de evaluación es muy útil cuando se tratan palabras con una distinción de sentidos muy fina, donde es posible anotar varios sentidos y que todos se consideren correctos. Una variante de la fórmula de la entropía cruzada sería utilizar la misma fórmula obviando el término logarítmico quedando tal y como muestra la Ecuación 3.3.

Universitat d'Alacant
Universidad de Alicante

$$Entropia\ cruzada\ adapt = \frac{1}{N}\sum_{i=1}^{N} Pr_S(sc_i|w_i, context_i) \quad (3.3)$$

Esta nueva medida puede ser utilizada por cualquier sistema, independientemente de que asigne probabilidades a cada posible sentido o no. En el caso de sistemas que sólo dan como resultado un único sentido, esta nueva fórmula es equivalente a la 3.1, que únicamente tenía en cuenta el sentido con mayor probabilidad.

3.1.2 Distancia semántica

Otro problema en el proceso de evaluación de sistemas de desambiguación automática, es que muchas medidas de evaluación no tienen en cuenta la distancia semántica entre sentidos a la hora de decidir si las palabras están anotadas correctamente. Esta situación es más evidente cuando hablamos de jerarquías de sentidos. La Tabla 3.2 muestra un ejemplo de agrupación de sentidos jerárquica para *"bank"* y su correspondiente matriz de distancia semántica.

I. Bank - Repository							
I.1 Financial bank		I.1a	I.1b	I.2	II.1	II.2	III
I.1a - the institution	I.1a	0	1	2	4	4	4
I.1b - the building	I.1b	1	0	2	4	4	4
I.2 General Supply/Reserve	I.2	2	2	0	4	4	4
II Bank - Geographical	II.1	4	4	4	0	1	4
II.1 Shoreline	II.2	4	4	4	1	0	4
II.2 Ridge/Embankment	III	4	4	4	4	4	0
III Bank - Array/Group/Row							

Tabla 3.2. Jerarquía de sentidos y matriz de distancia semántica para "bank"

Si en el proceso de anotación se produjera un error al anotar una palabra con el sentido correspondiente a un hermano dentro de la jerarquía de sentidos, la penalización debería ser más baja que si ese error se debiera a una anotación entre sentidos que no están relacionados de ninguna forma. La solución podría ser emplear una matriz que establezca la distancia semántica entre sen-

tidos de forma que cada celda $(sentido_1, sentido_2)$, contendrá el valor de la distancia entre sentidos. Cuanto más grande sea el valor de la celda, mayor será la distancia semántica entre los sentidos que representa. En la parte derecha de la Tabla 3.2 se muestra un ejemplo de matriz de distancia semántica para *"bank"*.

Una forma de evaluar los sistemas usando distancias semánticas es modificando la fórmula de la entropía cruzada, de forma que se trate de minimizar la distancia entre el sentido asignado (sa_i) y el sentido correcto (sc_i), sobre los N ejemplos, tal y como muestra la Ecuación 3.4.

$$Entropia\ cruzada\ dist = \frac{1}{N}\sum_{i=1}^{N} Distancia(sc_i, sa_i) \tag{3.4}$$

Además de tratar de minimizar la distancia semántica entre sentidos, también se podría medir la eficiencia de los sistemas penalizando las probabilidades asignadas a sentidos incorrectos según la Ecuación 3.5.

$$Eficiencia\ dist = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{S_i} Distancia(sc_i, s_j) \times Pr_S(s_j|w_i, context_i) \tag{3.5}$$

Donde para cada ejemplo i, se consideran todos los posibles sentidos (s_j) de la palabra w_i, midiendo las probabilidades que el sistema S ha asignado a sentidos incorrectos $Pr_S(s_j|w_i, context_i)$ por la distancia semántica de estos sentidos respecto al sentido correcto.

En (Melamed y Resnik (2000)) se propuso una variante de estas ideas para SENSEVAL, donde se utilizó el diccionario HECTOR (Atkins (1992)). La propuesta fue un esquema de distribución de probabilidades a través de los distintos niveles de la jerarquía de sentidos, donde además se podían incluir como respuestas válidas varios sentidos.

En SENSEVAL la evaluación de sistemas se ha hecho utilizando diferentes pautas, variando el nivel de granularidad (sentidos de

Universitat d'Alacant
Universidad de Alicante

bajo nivel frente a sentidos de alto nivel), etiquetación única de sentidos y etiquetación múltiple de sentidos. En la práctica, la mayoría de sistemas que han participado en SENSEVAL no obtienen un conjunto de probabilidades para todos los sentidos, por ello, la evaluación normalmente se realiza utilizando la variante de la entropía cruzada para un solo sentido según la Ecuación 3.3 comentada anteriormente.

3.2 Un marco común para la evaluación de sistemas

Como ya se ha comentado en el capítulo anterior, los sistemas de WSD supervisados y no supervisados tienen diferentes necesidades en cuanto a recursos necesarios para su evaluación. A pesar de que los sistemas no supervisados pueden ser evaluados con un corpus etiquetado como Semcor, que contiene una gran cantidad de palabras polisémicas, los sistemas supervisados necesitan de corpus más extensos para realizar su correspondiente entrenamiento y evaluación. Para establecer una guía de las necesidades tanto de sistemas supervisados como de sistemas no supervisados se ha desarrollado un protocolo utilizado en SENSEVAL con algunas modificaciones:

Universitat d'Alacant
Universidad de Alicante

1. Obtener un corpus extenso sin anotar (Por ejemplo, de N =1 billón de palabras).
2. Determinar el repositorio de sentidos a utilizar (WordNet, LDOCE) y sobre el cual se evaluarán los distintos sistemas.
3. Seleccionar un subconjunto de palabras $R < N$ (por ejemplo, 100 millones) en un corpus no anotado y proporcionarlo a los participantes.
4. Seleccionar un pequeño subconjunto de palabras $S < R < N$ (por ejemplo, 10 millones), como conjunto de test. Generar el test como sigue: (a) Seleccionar un conjunto M (por ejemplo, 100) palabras ambiguas. Estas palabras serán la base para la evaluación y no serán reveladas hasta el momento de distribuir el corpus de test. (b) Para cada una de las M palabras, anotar

todas las instancias de cada palabra en el corpus de test. (c) Para cada una de las M palabras, evaluar la anotación hecha por diferentes anotadores y establecer un acuerdo. (d) Para cada una de las M palabras, estudiar los casos en los que los anotadores no se ponen de acuerdo y tomar una decisión, por votación si fuera necesario.

5. Advertir a los participantes de no modificar el código de sus sistemas a partir de este punto.
6. Proporcionar a cada participante el corpus de test.
7. Evaluar cada sistema considerando todas las instancias de las M palabras anotadas para la evaluación. Comparar anotaciones de sentidos exactas, entropía cruzada, etc.
8. Almacenar el corpus utilizado como test para que sea empleado por sistemas de WSD supervisados. A partir de este punto podrán participar utilizando el corpus de test como corpus de entrenamiento.
9. Para la siguiente evaluación de sistemas ir al paso 3.

Concretamente en SENSEVAL se adoptaron algunos aspectos de este protocolo. El corpus del Paso 1 fueron 17 millones de palabras extraídas del British National Corpus[1], que hasta ahora ha ido incrementándose hasta 100 millones de palabras. El repositorio de sentidos seleccionado del Paso 2 fue la base de datos HECTOR (SENSEVAL-1. Como el proceso de evaluación incluía tanto sistemas supervisados como no supervisados, los corpus de entrenamiento incluyeron tanto instancias anotadas como no anotadas (al contrario que en el Paso 3), para un conjunto de 29 palabras ambiguas. Para el proceso de creación del corpus de test del Paso 4, se seleccionaron 34 palabras ambiguas distribuidas en 8448 instancias. Los participantes fueron advertidos de no modificar sus sistemas tal y como estaba especificado en el Paso 5.

La gran diferencia con el protocolo seguido en SENSEVAL fue con respecto a los Pasos 6 y 7, donde los participantes no sabían qué palabras iban a ser utilizadas para la evaluación. Por ello, se hicieron dos grupos, uno para evaluar a los sistemas que sólo desambiguaron las palabras del conjunto de test, y otro con los siste-

[1] http://www.natcorp.ox.ac.uk/

mas que desambiguaron todas las palabras con contenido semántico.

Los resultados obtenidos indicaron que los sistemas más eficientes eran aquellos que utilizaban corpus de entrenamiento para aprender clasificadores especialmente diseñados para las palabras del corpus de test.

En el Capítulo 6 se describen en profundidad todas las ediciones de SENSEVAL hasta la actualidad y algunos de los sistemas más relevantes en la tarea de WSD.

Universitat d'Alacant
Universidad de Alicante

Capítulo 4

Recursos

A continuación se van a describir los recursos utilizados como base para el desarrollo de nuestros métodos de desambiguación y para la creación del nuevo recurso léxico Dominios Relevantes. La elección de estos recursos se debe al conjunto de características (relaciones semánticas entre palabras, conexión con varios idiomas, adquisición de conocimiento a través de relaciones entre palabras, etc) que los hacen idóneos para la tarea de WSD.

4.1 WordNet

WordNet (Fellbaum (1998)) fue concebido como un diccionario electrónico siguiendo principios psicolingüísticos. Su contenido se organiza mediante una base de datos léxica donde se agrupan conjuntos de palabras (nombres, verbos, adjetivos y adverbios) en grupos de sinónimos llamados synsets: un synset se codifica como un número único de ocho dígitos. Dentro de la base de datos, cada synset representa un concepto distinto y entre cada uno de ellos existen conexiones que expresan relaciones semánticas, conceptuales o léxicas. El resultado de este conjunto de conexiones es una extensa red navegable que proporciona un gran número de inter-relaciones entre palabras. Entre este conjunto de relaciones encontramos las siguientes:

- **Sinonimia.** Dentro de la misma categoría sintáctica (nombre, verbo, adjetivo o adverbio), son sinónimas aquellas palabras que pueden sustituirse dentro de un contexto determinado sin alterar su significado. Por ejemplo, en las frases: "Me pasó una hoja en blanco" y "Me pasó un folio en blanco". Las palabras "hoja" y "folio" son sinónimas porque al sustituir una por la otra no se altera el significado de la frase.

  Sinónimos para *"bank#1"* { *depository financial institution#1*, *banking concern#1*, *banking company#1*, *financial institution#1*, *financial organization#1*, *financial organisation#1*

- **Antonimia.** Son antónimas aquellas palabras con significados opuestos. Por ejemplo: seco/mojado, subir/bajar, avanzar/retroceder ...

  Antónimos para *"clean#1"* { *dirty#1*, *soil#1*, *begrime#1*, *grime#1*, *colly#1*, *bemire#1*

- **Hiponimia.** Mientras que la sinonimia y la antonimia son relaciones léxicas entre palabras, la hiponimia es una relación entre los significados de las palabras. Estas relaciones se dan únicamente para los nombres. Por ejemplo: "arce" es un hipónimo de "árbol" y "árbol" es un hipónimo de "planta". Este tipo de relación se conoce también con el nombre de "IS_A". Se entiende que "X" es un hipónimo de "Y" si **"X es un (tipo de) Y"**.

Hipónimos para *"cat#1"* { *domestic cat#1*, *house cat#1*, *Felis domesticus#1*, *Felis catus#1*, *wildcat#3*

Hiperonimia. Esta relación se define como la inversa de la hiponimia. Es decir, Y es un hiperónimo de X si **"X es un (tipo de) Y"**.

Hiperónimos para *"cat#1"* { *feline#1*, *felid#1*, *carnivore#1*, *placental#1*, *placental mammal#1*, *eutherian#1*, *eutherian mammal#1*, *mammal#1*, *mammalian#1*, *vertebrate#1*

- **Meronimia.** La sinonimia, antonimia, hiponimia e hiperonimia son relaciones aplicadas comúnmente. Otra relación semántica es la meronimia, identificada como un tipo de vínculo "HAS_A". Una palabra X es merónima de Y si **" X es una parte de Y "**. Por ejemplo, párpado, retina o córnea son merónimos de ojo, porque todos son partes del ojo.

Merónimos para *"body#1"*
- *articulatory system#1*
- *digestive system#1*
- *gastrointestinal system#1*
- *endocrine system#1*
- *lymphatic system#1*
- *musculoskeletal system#1*
- *sensory system#2*
- *trunk#3*

- **Holonimia.** Esta relación se define como la inversa de la meronimia. Es decir, Y es un holónimo de X si **" X es una parte de Y "**. Por ejemplo, "casa" es un holónimo de "dormitorio", "comedor", "cocina", etc.

  Holónimos para *"eye#1"*
  - *visual system#1*
  - *face#1*
  - *human face1#1*

- **Troponimia.** La troponimia relaciona verbos y es el equivalente de la relación de hiponimia para los nombres.

  Tropónimos para *"eat#1"*
  - *wash down#1*
  - *gluttonize#1*
  - *gluttonise#1*
  - *fress#1*
  - *wolf#1*
  - *slurp#1*
  - *fare#2*

- **Entailment.** En esta relación un término implica al otro. Por ejemplo, divorcio/matrimonio.

Universitat d'Alacant
Universidad de Alicante

$$\text{Entailment para “eat\#1”} \begin{cases} \textit{chew\#1} \\ \textit{masticate\#2} \\ \textit{manducate\#1} \\ \textit{jaw\#3} \\ \textit{swallow\#1} \\ \textit{get down\#4} \end{cases}$$

Además de distinguir mediante synsets los significados de cada término, WordNet establece una relación de orden entre los diferentes sentidos de las palabras, de acuerdo a su frecuencia de aparición. De esta forma, para *“plant”* en la versión 2.1 existen cuatro significados diferentes:

1. **{*03912097*}** *plant#1, works#1, industrial plant#1 – (buildings for carrying on industrial labor; “they built a large plant to manufacture automobiles”)*
2. **{*00016858*}** *plant#2, flora#2, plant life#1 – (a living organism lacking the power of locomotion)*
3. **{*05831211*}** *plant#3 – (something planted secretly for discovery by another; “the police used a plant to trick the thieves”; “he claimed that the evidence against him was a plant”)*
4. **{*10282477*}** *plant#4 – (an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience)*

Universidad de Alicante

Cada concepto asociado al término *“plant”* además de tener asociado su synset, también tiene asociado un número de sentido: *plant#1, plant#2, plant#3* y *plant#4*, estos sentidos indican la frecuencia de aparición de cada concepto, siendo plant#1 el más frecuente.

Estrechamente vinculada a cada synset existe una definición o glosa que define el concepto representado por el sentido específico de cada término. Así para *plant#1* su glosa asociada es *“buildings for carrying on industrial labor”*. Además, en la mayoría de términos (synsets), se sitúa la palabra en un contexto (oración),

ejemplificando de esta forma su uso: *"they built a large **plant** to manufacture automobiles"*.

Los conceptos situados en lo alto de la jerarquía de WordNet de los que derivan el resto de conceptos, son los mostrados en la Tabla 4.1.

Concepto	Definición
entity	that which is perceived or known or inferred to have its own distinct existence (living or nonliving)
psychological feature	a feature of the mental life of a living organism
abstraction	a general concept formed by extracting common features from specific examples
state	the way something is with respect to its main attributes; "the current state of knowledge"; "his state of health"; "in a weak financial state"
event	something that happens at a given place and time
act, human action, human activity	something that people do or cause to happen
group, grouping	any number of entities (members) considered as a unit
possession	anything owned or possessed
phenomenon	any state or process known through the senses rather than by intuition or reasoning

Tabla 4.1. Conceptos en la cima de la jerarquía de WordNet

Universitat d'Alacant

En la Figura 4.1 tenemos una representación de la red semántica para la palabra *"aircraft"* con sentido 1. En esta imagen se puede observar la extensión de las diferentes dimensiones que puede tener un concepto en WordNet.

Otro ejemplo de las relaciones semánticas existentes en WordNet lo encontramos en la Figura 4.2, donde se muestra un extracto de la relaciones existentes para *"bank#1"*.

Como se puede apreciar en la Figura 4.2, para *bank#1* existen una serie de sinónimos: *banking_company, banking_concern* y *depository_finantial_institution*. Así como una serie de hipónimos representados mediante flechas de color verde, merónimos repre-

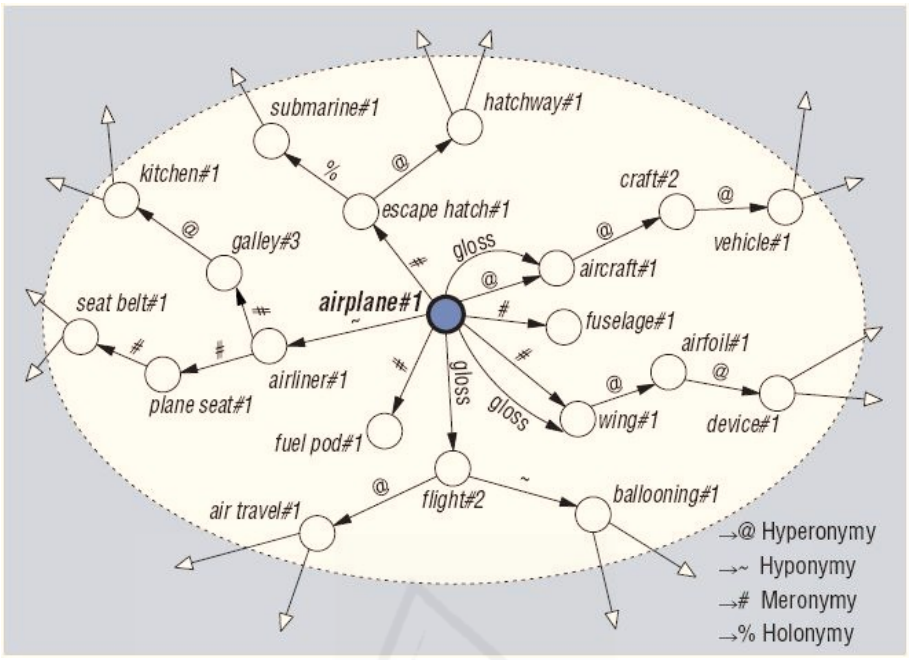


Figura 4.1. Red semántica para airplane#1

sentados mediante flechas de color amarillo y holónimos representados mediante flechas de color morado.

Una representación más detallada de esta versión gráfica de las relaciones existentes para *bank#1* se muestra en la Tabla 4.2.

4.2 WordNet Domains

WordNet Domains extiende la información proporcionada por WordNet mediante la inclusión de *"Subject Field Codes" (SFC)*, es decir, conjuntos de palabras relevantes para un dominio específico. Una representación de este tipo de información la encontramos en las etiquetas de campo semántico, de uso común en todo tipo de diccionarios, por ejemplo: MATEMÁTICAS, BOTÁNICA, etc. Por un lado, estas etiquetas clarifican a qué contexto se refiere la definición que sigue, por ejemplo, la palabra *"anillo"*, pertenece a diferentes contextos, tales como, ARQUITECTURA "Cornisa circular u ovalada", BOTÁNICA "Cada uno de los círculos leñosos

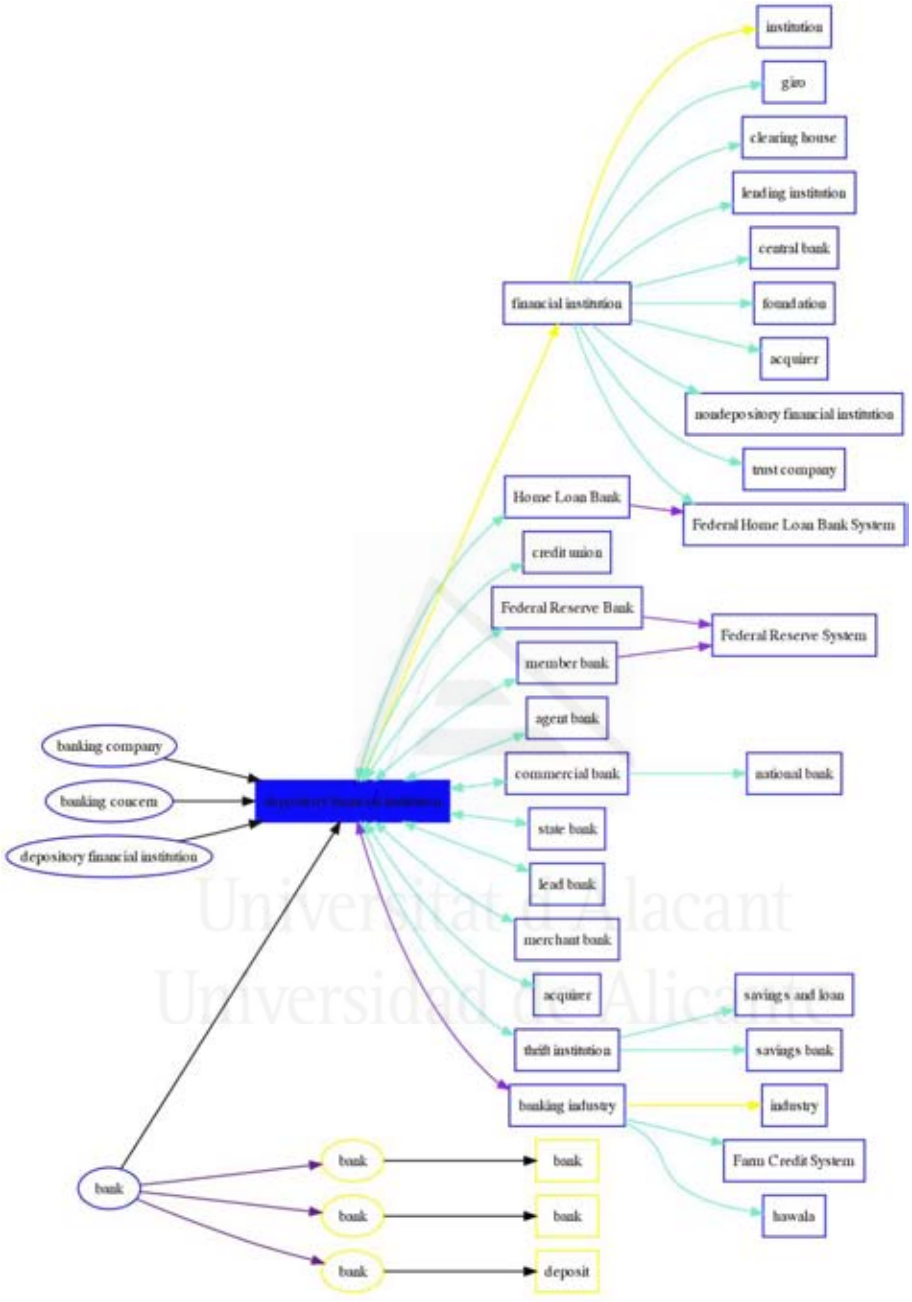


Figura 4.2. Relaciones semánticas para bank#1

Relaciones	Palabras	Glosa
Sinónimos	Depository_financial _institution Bank Banking_concern Banking_company	A financial institution that accepts deposits and channels the money into lending activities; "he cashed a check at the bank"; "that bank holds the mortgage on my home"
Hipónimos	Credit_union	A cooperative depository financial institution whose members can obtain loans from their combined savings
	Home_Loan_Bank	One of 11 regional banks that monitor and make short-term credit advances to thrift institutions in their region
	Federal_Reserve_Bank	One of 12 regional banks that monitor and act as depositories for banks in their region
	Member_bank	A bank that is a member of the Federal Reserve System
	Agent_bank	A bank that acts as an agent for a foreign bank
	Commercial_bank	A financial institution that accepts demand deposits and makes loans and provides other services for the public
	...	...
Merónimos	Financial_institution	an institution (public or private) that collects funds (from the public or other institutions) and invests them in financial assets)
Holónimos	Banking_industry	banks collectively

Tabla 4.2. Relaciones existentes para bank#1

Universitat d'Alacant
Universidad de Alicante

concéntricos que forman el tronco de un árbol" y MATEMÁTICAS "Conjunto de elementos entre los que se definen dos reglas de composición". Por otro lado, estas etiquetas permiten la búsqueda rápida de la acepción deseada, por ejemplo, si buscamos el significado de *"disco"* dentro del contexto de la INFORMÁTICA no es necesario ir leyendo todas las acepciones una por una hasta dar con la que deseamos, simplemente basta con mirar la etiqueta del campo semántico que precede a cada definición hasta dar con la que nos interesa, en este caso, INFORMÁTICA.

Dada su utilidad, los *SFC* ya han sido usados en Lingüística y en Lexicografía para marcar los usos técnicos de las palabras. Aunque ésta es una información muy útil para establecer una

discriminación de sentidos, en los diccionarios generalmente se emplea sólo para una pequeña parte del léxico. Un ejemplo lo tenemos en la entrada para la palabra *"bolsa"*, (Tabla 4.3), en el Diccionario de la Lengua de la Real Academia Española (R.A.E).

Definiciones de la palabra Bolsa en el R.A.E.
1. f. Especie de talega o saco de tela u otro material, que sirve para llevar o guardar algo.
2. f. Saco pequeño de cuero en que se echa dinero, y que se ata o cierra.
3. f. Recipiente de material resistente para guardar, en viajes o traslados, ropa u otras cosas, y que se puede llevar a mano o colgado del hombro. *Bolsa de deporte.*
4. f. **folgo.**
5. f. Arruga que hace un vestido cuando viene ancho o no ajusta bien al cuerpo, o la que forman dos telas cosidas cuando una es más larga o ha dado de sí más que la otra.
6. f. Abultamiento de la piel debajo de los ojos.
7. f. Acumulación de un fluido en un determinado lugar.
8. f. Caudal o dinero de una persona. *A Juan se le acabó la bolsa.*
9. f. Pieza de estera en forma de saco, que pende entre los varales del carro o galera, y debajo de la zaga de los coches o calesas, para colocar efectos.
10. f. Taleguilla de tafetán o moaré negro con una cinta en la parte superior que usaban los hombres para llevar recogido el pelo.
11. f. *Dep.* Premio en metálico que recibe el ganador de un combate de boxeo.
12. f. *Dep.* Cantidad que se ofrece a quien participa en otras competiciones.
13. f. *Ingen.* Parte de un criadero donde el mineral está reunido con mayor abundancia.
14. f. *Med.* Cavidad llena de pus, linfa, etc.
15. f. *Mil.* Situación en que queda un ejército o una parte de él al ser completamente rodeado por las fuerzas enemigas.
16. f. *Am. Cen.* y *Méx.* Bolsillo de las prendas de vestir.
17. f. pl. Cavidades del escroto en las cuales se alojan los testículos.
18. f. pl. u. c. sing. m. vulg. *Ven.* Persona imbécil, lerda.

Tabla 4.3. Definiciones para la palabra "bolsa" del RAE

Con el fin incorporar la información de las etiquetas semánticas a WordNet se construyó un nuevo recurso llamado WordNet Domains (Magnini y Cavaglia (2000)). Mediante WordNet Domains se pretendía mejorar la distinción de los sentidos en WordNet, agrupando en muchos casos distintos sentidos bajo un mismo dominio o categoría semántica.

Mediante WordNet Domains, se intenta extender la cobertura de las etiquetas de dominio, dentro de una base de datos léxica ya existente: WordNet. En WordNet Domains los synsets de WordNet han sido anotados mediante un proceso semiautomático con una o varias etiquetas de dominio, seleccionadas de entre un conjunto de 200 etiquetas organizadas jerárquicamente.

La anotación de WordNet mediante *SFC*'s viene motivada por:

- **Crear nuevas relaciones entre palabras.** Mediante las etiquetas de dominio se pueden establecer relaciones entre palabras que pertenecen a distintas categorías, ya que, por ejemplo, en WordNet1.6 [1], no encontramos relaciones entre nombres y verbos.
- **Anotar a nivel semántico.** Debido a que los dominios se asocian a synsets, la anotación se realiza a nivel semántico y no a nivel de palabra.
- **Obtener recursos multilingües.** Los *SFC*'s son básicamente independientes del lenguaje, por lo que se pueden incluir en recursos multilingües tales como, EuroWordNet (Vossen (1998)).

La información aportada por los *SFC*'s es complementaria a la información que tiene WordNet. La primera característica que añaden a WordNet es que dentro de un mismo dominio pueden incluirse synsets que pertenecen a diferentes categorías sintácticas.

Por ejemplo, con el dominio MUSIC se han anotado palabras pertenecientes a diferentes categorías sintácticas, tal como se muestra en la Tabla 4.4.

Una segunda característica que aportan los dominios a WordNet es que dentro de un mismo dominio pueden aparecer sentidos de palabras pertenecientes a diferentes subjerarquías de WordNet, es decir, descendientes de diferentes raíces o de diferentes ficheros lexicográficos. Por ejemplo, el dominio SPORT contiene sentidos como *athlete#1*, que deriva de *life_form#1*, *game_equipment#1*, derivado de *physical_object#1*, *sport#1* derivado de *act#2* y *playing_field#1*, derivado de *location#1*.

[1] En la versión 2.0. ya se pueden establecer este tipo de relaciones

Dominio Music		
Categoría	**Palabras**	**Glosas**
Nombres	album#1	one or more phonograph records or tape recordings issued together
	band#2	instrumentalists not including string players
	bar#3	notation for a repeating pattern of musical beats; written followed by a vertical bar
Verbos	compose#2	write music; "Beethoven composed nine symphonies"
	drum#2	play the drums
	modulate#1	change the key of, in music; "modulate the melody"
Adjetivos	bowed#2	(music) of a stringed instrument; sounded by stroking with a bow
	chromatic#1	(music) based on a scale consisting of 12 semitones; "a chromatic scale"
Adverbios	fugally#1	(music) in a fugal style
	presto#2	(music) at a very fast tempo (faster than allegro)

Tabla 4.4. Relaciones entre diferentes categorías sintácticas mediante el uso de dominios.

La tercera y última característica que añaden los dominios a WordNet, es la posibilidad de reducir el nivel de polisemia de las palabras, es decir, dentro de un mismo dominio se pueden agrupar diferentes sentidos pertenecientes a una misma palabra. Por ejemplo, los dominios asociados a la palabra *"man"*, que en WordNet tiene 10 sentidos, son los que se muestran en la Tabla 4.5.

Si se anotan los sentidos utilizando dominios tal y como aparecen en la Tabla 4.5, se puede reducir el nivel de polisemia de 10 sentidos a 4 sentidos, agrupando aquellos sentidos que pertenecen a un mismo dominio. En este caso, se agruparían dentro de un único sentido todos aquellos conceptos pertenecientes al dominio PERSON:

man#1,3,5,6,7,8,9 ⇒ PERSON
man#2 ⇒ MILITARY
man#4 ⇒ FACTOTUM
man#10 ⇒ PLAY

Palabra	Dominio	Glosa
man#1	person	an adult male person (as opposed to a woman); "there were two women and six men on the bus"
man#2	military	someone who serves in the armed forces; "two men stood sentry duty"
man#3	person	the generic use of the word to refer to any human being; "it was every man for himself"
man#4	factotum	all of the inhabitants of the earth; "all the world loves a lover"
man#5	biology, person	any living or extinct member of the family Hominidae
man#6	person	a male subordinate; "the chief stationed two men outside the building"; "he awaited word from his man in Havana"
man#7	person	an adult male person who has a manly character (virile and courageous competent); "the army will make a man of you"
man#8	person	(informal) a male person who plays a significant role (husband or lover or boyfriend) in the life of a particular woman; "she takes good care of her man"
man#9	person	a manservant who acts as a personal attendant to his employer; "Jeeves was Bertie Wooster's man"
man#10	play	a small object used in playing certain board games; "he taught me to set up the men on the chess board"; "he sacrificed a piece to get a strategic advantage"

Tabla 4.5. Reducción de la polisemia mediante el uso de dominios

Universitat d'Alacant
Universidad de Alicante

Los dominios se han estructurado desde dos puntos de vista diferentes: jerárquicamente y semánticamente.

- **Estructuración jerárquica.** En la jerarquía de dominios encontramos diferentes niveles de especificación. Por ejemplo, dentro del nivel 1 podemos encontrar dominios de tipo Botany, Linguistics, History y Religion. Sin embargo, dentro del nivel 2 encontramos dominios de tipo Building_Industry, Dentistry, Football y Photography. Cuanto más profundizamos es los niveles de la jerarquía, mayor es el nivel de especialización de los dominios.
  En la figura 4.3 se muestra un pequeño fragmento de la jerarquía de WordNet Domains.

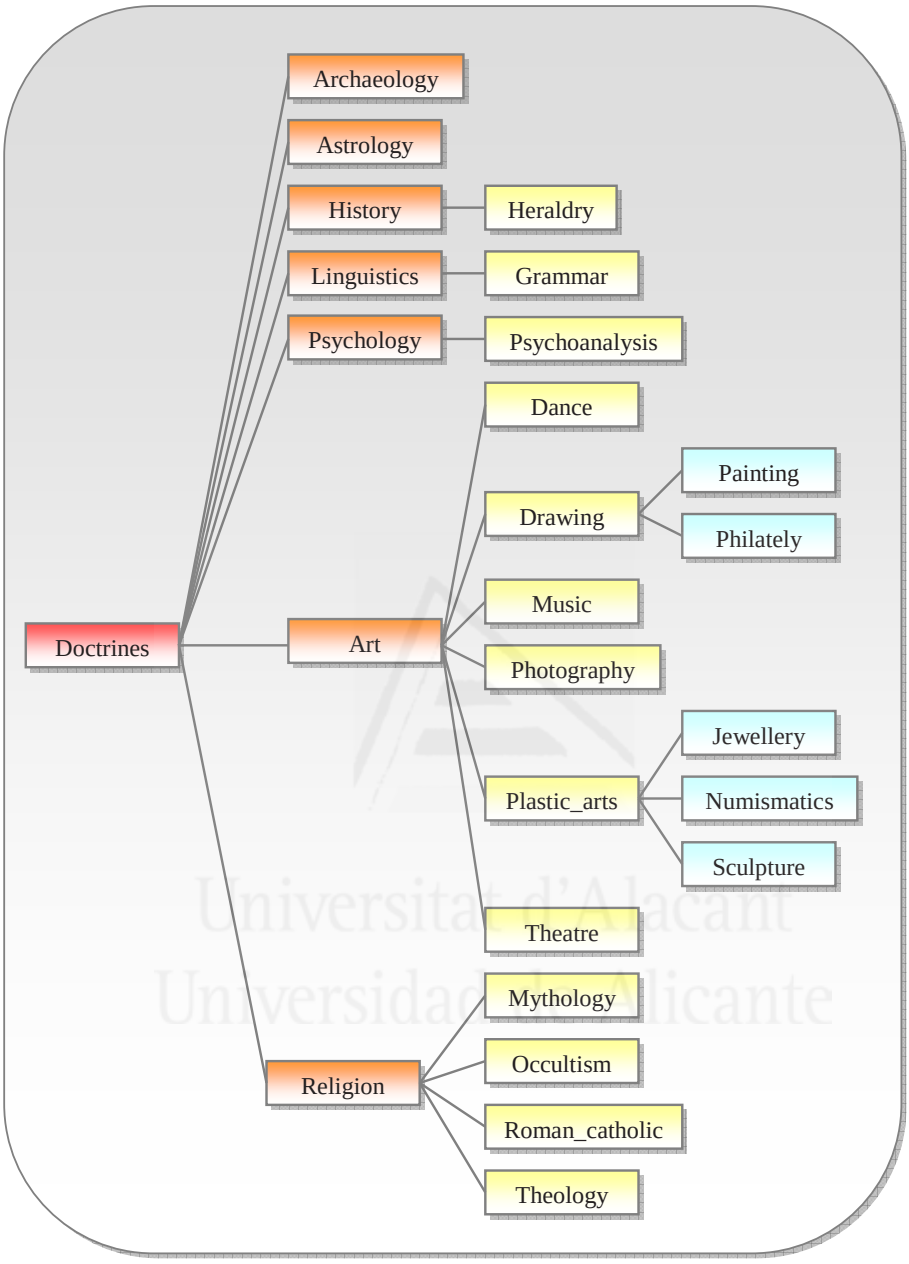


Figura 4.3. Jerarquía de WordNet Domains

- **Estructuración semántica.** Además de la estructuración jerárquica, los dominios se organizan en familias. Una familia es un conjunto de dominios semánticamente relacionados entre los que no existen relaciones de inclusión. Mientras que la organización jerárquica es fija, la organización por familias se puede reorganizar permitiendo la creación de nuevas relaciones interdisciplinarias.

Entre el conjunto de SFC's encontramos una etiqueta de dominio denominada FACTOTUM. Este dominio se ha creado exclusivamente para englobar dos tipos de synsets:

- **Synsets genéricos.** Estos synsets son aquellos que difícilmente se pueden clasificar dentro de algún dominio en particular. Por ejemplo: act#2: something that people do or cause to happen act#5: a manifestation of insincerity; "he put on quite an act for her benefit"
- **Stop senses.** Son aquellos synsets que aparecen frecuentemente en diferentes contextos, tales como números, días de la semana, colores, etc. Estos synsets pertenecen normalmente a palabras monosémicas.

El proceso de anotación de los synsets de WordNet1.6 mediante *SFC's* se divide en tres pasos:

- **Paso 1.** Un número reducido de synsets pertenecientes a los niveles más altos de la jerarquía de WordNet son anotados manualmente mediante los SFC's.
- **Paso 2.** A partir de la anotación obtenida en el paso 1, se ejecuta un proceso automático que explota las relaciones de WordNet (hiponimia, troponimia, meronimia, antonimia, etc), para extender la anotación manual a todos aquellos synsets alcanzables.
- **Paso 3.** El último paso es realizar la evaluación de los resultados obtenidos por el proceso automático. Las anotaciones erróneas son detectadas, se corrigen y se vuelve a lanzar el proceso del paso 2 a partir de los nuevos valores.

Universitat d'Alacant
Universidad de Alicante

4.3 Extended WordNet

Extended WordNet (Harabagiu et al. (1999)) es un nuevo recurso léxico creado en la Universidad de Texas que trata de mejorar la información proporcionada por WordNet en sus distintas versiones, agregando información semántica a las glosas. Esta nueva información se extrae únicamente de la parte de la definición de las glosas, descartando los ejemplos y las aclaraciones entre paréntesis que puedan aparecer. Por ejemplo, pensemos en la glosa de la palabra *"man"*: *an adult male person who has a manly character (virile and courageous competent); "the army will make a man of you"*. En este caso, únicamente se tendría en cuenta la información proporcionada en la definición *"an adult male person who has a manly character"*, omitiendo la aclaración entre paréntesis y la frase de ejemplo para este sentido.

Para ampliar y mejorar la información proporcionada por WordNet se realizan tres tipos de análisis diferentes sobre las glosas:

- **Análisis sintáctico.** Para obtener el análisis sintáctico de las glosas se ha utilizado una versión mejorada del etiquetador de Brill (Brill (1995)). Esta nueva versión ha sido entrenada sobre WordNet. El resultado obtenido (palabras con su categoría sintáctica, género, número,...), se utiliza como entrada para dos tipos distintos de analizadores sintácticos. Estos analizadores se integran en un esquema de votación para tratar de mejorar la calidad de los resultados. Aunque este análisis sintáctico, se podría haber aplicado directamente sobre las glosas de WordNet sin un preproceso inicial, se consideró necesario un tratamiento previo de las glosas para obtener unos resultados más precisos. Este tratamiento previo consiste en extender el contenido de las glosas de la siguiente forma:
  - **Adverbios.** Las glosas pertenecientes a adverbios se extienden añadiendo (el adverbio + *is*) al principio de la glosa y un punto al final de la definición. Por ejemplo, para el adverbio *"entirely"* su glosa quedaría como sigue: ***entirely is*** without any others being included or involved**.**

- **Adjetivos.** Las glosas pertenecientes a adjetivos se extienden añadiendo (el adjetivo + *is something*) al principio de la glosa y un punto al final de la definición. Por ejemplo, para el adjetivo *"infinite"* su glosa quedaría como sigue: ***infinite is something*** total and all-embracing**.**
- **Verbos.** Las glosas pertenecientes a verbos se extienden añadiendo (*to* + el verbo + *is to*) al principio de la glosa y un punto al final de la definición. Por ejemplo, para el verbo *"hiccup"* su glosa quedaría como sigue: ***to hiccup is to*** breathe spasmodically , and make a sound**.**
- **Nombres.** Las glosas pertenecientes a nombres se extienden añadiendo (el nombre + *is*) al principio de la glosa y un punto al final de la definición. Por ejemplo, para el nombre *"space"* su glosa quedaría como sigue: ***space is*** the unlimited 3-dimensional expanse in which everything is located**.**

▪ **Análisis lógico.** Muchas aplicaciones necesitan hacer uso de la información pragmática contenida en los textos. Debido a esta necesidad, Extended WordNet incorpora información lógica a las glosas. Utilizando las glosas conceptuales originales de WordNet, éstas se transforman en su forma lógica correspondiente. La forma lógica obtenida es un paso intermedio entre el análisis sintáctico y una forma semántica. Es decir, esta transformación codifica las relaciones sintácticas siguientes: 1) sujetos sintácticos, 2) objetos sintácticos, 3) enlaces preposicionales, 4) nominales complejos y 5) adjuntos adjetivales y adverbiales.

▪ **Análisis semántico.** En WordNet existen más de 115000 glosas. Estas definiciones (glosas) se encuentran repartidas de la siguiente forma: unas 79000 están asociadas a nombres, alrededor de 13000 a verbos, sobre 18000 a adjetivos y por último unas 3500 a adverbios. En la tarea de análisis semántico se han utilizado dos tipos de anotación: automática y manual. En la anotación automática han intervenido dos sistemas: uno diseñado de forma específica para desambiguar las glosas de WordNet, llamado XWN_WSD y un sistema propio para desambiguar texto libre. Para decidir el sentido asociado a cada una de las palabras de la glosa, se ha empleado un sistema

Universitat d'Alacant
Universidad de Alicante

de votación entre los dos métodos automáticos. De forma que la precisión estimada cuando los dos métodos etiquetan una palabra con el mismo sentido es del 90 %. Existen tres categorías de anotación semántica ordenadas según su fiabilidad: "GOLD", "SILVER" y "NORMAL". Cuando el sentido seleccionado para una palabra se etiqueta como "GOLD" quiere decir esa anotación se ha comprobado de forma manual. Cuando un sentido aparece con la etiqueta "SILVER" significa que ha sido etiquetado de forma consensuada por los dos métodos automáticos de desambiguación. Y por último cuando un sentido aparece junto a la etiqueta "NORMAL" quiere decir que se ha seleccionado el sentido proporcionado por el sistema automático XWN_WSD. Cabe considerar que los verbos "to be" y "to have" se han tratado de forma especial y no se han desambiguado de forma automática. El procedimiento seguido para obtener la desambiguación de las glosas de WordNet consta de dos fases:

1. El primer paso consta de la realización de un preproceso sobre las glosas de WordNet para separar la parte de la definición de la parte de ejemplos. También se realiza la tokenización, el análisis sintáctico usando el etiquetador de Brill y la identificación de conceptos compuestos.
2. El segundo paso, una vez realizado el preproceso, es asignar a cada palabra de la glosa su correspondiente sentido. En esta fase, se utiliza la categoría sintáctica obtenida en la fase previa para establecer el sentido correspondiente, ya que, una misma palabra puede actuar como nombre, verbo, adjetivo o adverbio, y por tanto, puede adoptar distintos significados dependiendo de su categoría.

   Para poder realizar correctamente el proceso de desambiguación se han empleado distintas heurísticas (método XWN_WSD):

   - Palabras monosémicas: Se identifican y se etiquetan con el sentido 1.
   - Misma familia jerárquica: Se identifican aquellas palabras de la glosa que pertenecen a la misma jerarquía que el synset de la glosa.

- Paralelismo léxico: Se identifican aquellas palabras que pertenecen a la misma categoría sintáctica y que están separadas por comas o conjunciones. Cuando es posible, se trata de seleccionar los sentidos que pertenecen a la misma jerarquía.
- Búsqueda en Semcor: Dada una palabra de la glosa se forman dos parejas: palabra-palabra_siguiente_de_la_glosa palabra_anterior_de_la_glosa-palabra. Estos dos pares de palabras se buscan en el corpus de Semcor (G. Miller y Bunker (1993)). Si para estos pares de palabras el sentido asignado es siempre el mismo y el número de ocurrencias supera un cierto umbral, se le asigna ese sentido.
- Dominio asociado: Cada glosa de WordNet se ha anotado con un dominio (Magnini y Strapparava (2000)). Si la palabra de la glosa tiene algún sentido anotado con el mismo dominio que el de la glosa, se selecciona ese sentido.
- ...

Mediante estas heurísticas se han desambiguado el 64 % de las palabras de WordNet con un 75 % de precisión. El resto de las palabras se han etiquetado con el sentido#1 (el más frecuente). Además del método anterior, se ha utilizado otro método automático para el proceso de desambiguación, el cual, utiliza como base texto libre. Para su aplicación fue necesaria la transformación de las glosas en oraciones completas. De esta forma, se obtuvo una cobertura de un 100 % y una precisión del 70 %. Las palabras que se etiquetaron con el mismo sentido por los dos sistemas obtuvieron un 90 % de precisión.

En la Tabla 4.6 aparece el resultado obtenido tras realizar los tres tipos de análisis sobre la glosa asociada al adjetivo ***"excellent"***.

1. excellent, first-class, fantabulous – (of the highest quality; “made an excellent speech”; “the school has excellent teachers”; “a first-class mind”)

(TOP (S (NP (JJ excellent))
(VP (VBZ is)
(NP (NP (NN something))
(PP (IN of)
(NP (DT the) (JJS highest) (NN quality)))))
(. . .)))

excellent: JJ(x1) ⟶ of: IN(x1, x2) highest: JJ(x1)
quality: NN(x1)

<wf pos=“IN”> of < /wf>
<wf pos=“DT”> the < /wf>
<wf pos=“JJS” lemma=“high” quality=“silver” wnsn=“1”> highest < /wf>
<wf pos=“NN” lemma=“quality” quality=“normal” wnsn=“2”> quality < /wf>

Tabla 4.6. Excellent#1: Análisis sintáctico, formas lógicas y anotación semántica.

4.4 SUMO (Suggested Upper Merged Ontology)

La ontología SUMO (Suggested Upper Merged Ontology) es una ontología de nivel superior (Niles y Pease (2001)). Esta ontología proporciona definiciones para términos de propósito general y puede actuar como base para ontologías de dominios más específicos. SUMO fue creada a partir de la combinación de diferentes contenidos ontológicos en una única estructura cohesiva y actualmente existen alrededor de 1000 términos y 4000 aserciones. Los contenidos a partir de los cuales se obtuvo SUMO proceden de: Ontolingua[2], John Sowa’s upper level ontology[3] y las ontologías desarrolladas por ITBM-CNR (Unrestricted-Time, Representation, Anatomy, Biologic-Functions, and Biologic-Substances). El lenguaje de representación utilizado es una versión de KIF (Knowledge Interchange Format) (Genesereth (1991)), llamada SUO-KIF.

[2] http://www.ksl.stanford.edu/software/ontolingua/
[3] http://www.jfsowa.com/ontology/

El proceso de creación de esta ontología consta de varios pasos. Primero, se identificaron todos los contenidos ontológicos de alto nivel. Estos contenidos incluían las librerías de ontologías disponibles en el servidor Ontolingua y ITBM-CNR, la ontología de John Sowa, la ontología de Russell y Norvig (Russell y Norvig (1995)), los axiomas temporales de James Allen (Allen (1984)) y otras representaciones. Una vez extraído todo el contenido relevante se transformó al lenguaje SUO-KIF. Una vez realizada la traducción, el paso más complicado fue la creación de una única ontología que combinara todo el contenido recopilado. Para llevar a término este proceso, en primer lugar, se dividieron los conceptos en dos grupos: conceptos de alto nivel y conceptos de bajo nivel. En el primer grupo, se mantuvo la ontología de John Sowa y la ontología de Russell y Norvig. En el segundo grupo se incluyó el resto. Tras la división las dos ontologías de alto nivel ambas se combinaron para obtener una única estructura conceptual. El resto del contenido de las clases de bajo nivel fue añadido tras la combinación. La forma de incluir las clases de bajo nivel y los problemas a los que se tuvo que hacer frente están descritos en (Niles y Pease (2001)).

Para comprender la estructura y el contenido de SUMO podemos extraer los conceptos de más alto nivel, tal y como muestra la Figura 4.4.

Al igual que en la mayoría de jerarquías el concepto de más alto nivel es *"entity"* y bajo este concepto se encuentran *"physical"* y *"abstract"*.

En la Figura 4.6 se muestra un ejemplo de la jerarquía de conceptos y relaciones existentes en SUMO para bank#1.

En esta representación gráfica existe un código de colores asociado a las distintas clases de elementos y sus relaciones según la Figura 4.5.

Universitat d'Alacant
Universidad de Alicante

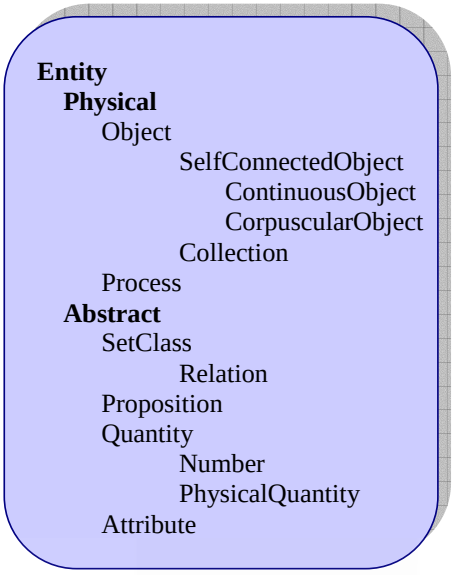


Figura 4.4. Conceptos de alto nivel en SUMO

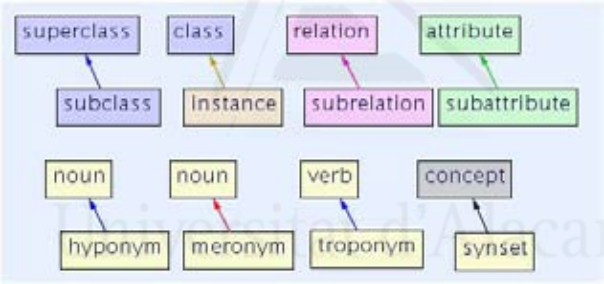


Figura 4.5. Código de colores en la representación gráfica de SUMO

Universitat d'Alacant
Universidad de Alicante

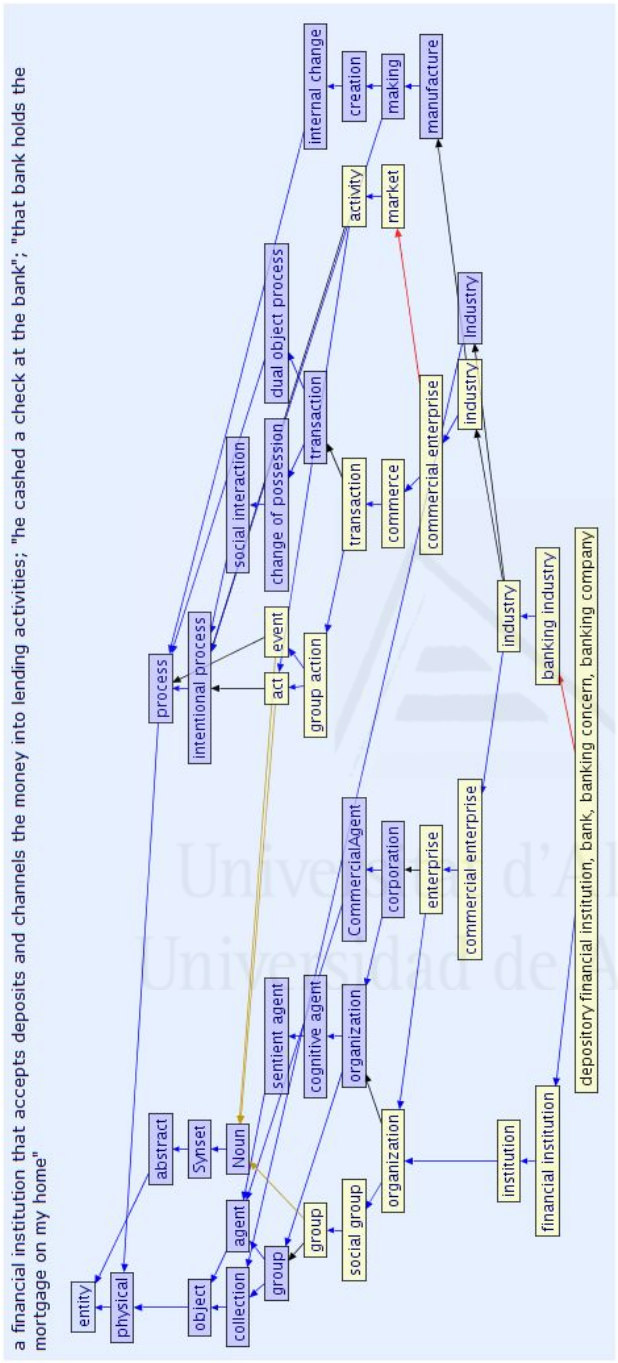


Figura 4.6. Jerarquía SUMO para bank#1

La ontología SUMO también se ha enlazado con los synsets de WordNet (Niles y Pease (2003)). De la misma forma que los Subject Field Codes se integraron en WordNet Domains, los diferentes conceptos de la ontología de SUMO se han enlazado con los synsets de WordNet.

En el proceso de anotación de WordNet con la ontología SUMO se han utilizado tres tipos de relaciones: sinonimia, hiperonimia e instanciación. A continuación se muestra un ejemplo para cada una de este tipo de relaciones:

Sinonimia. En caso de utilizar relaciones de sinonimia, veamos el ejemplo de la palabra *"plant"* cuyo synset en WordNet 1.6 es 00008864.

00008864 03 n 03 plant 0 flora 0 plant_life 0 027 @ . . . — a living organism lacking the power of locomotion

En este caso, el synset 00008864 es sinónimo del concepto de la ontología SUMO PLANT. Por tanto, la entrada en WordNet se amplía de la siguiente forma:

00008864 03 n 03 plant 0 flora 0 plant_life 0 027 @ . . . — a living organism lacking the power of locomotion &%Plant=

El prefijo "&%" indica que el concepto se ha obtenido a partir de SUMO y el signo "=" indica que el mapeo ha utilizado una relación de sinonimia.

Hiperonimia. En caso de que un synset de WordNet no tenga una correspondencia exacta con un concepto de la ontología de SUMO, se utiliza la relación de hiperonimia. Supongamos que tenemos la palabra *"Christian_Science"*, cuya entrada en WordNet 1.6 es la siguiente:

04719796 09 n 01 Christian_Science 0 001 @ 04718274 n 0000 — religious system based on teachings of Mary Baker Eddy emphasizing spiritual healing

En este caso, no existe un concepto específico para *"Chris-*

Universitat d'Alacant
Universidad de Alicante

tian_Science", sin embargo, la ontología contiene conceptos más generales, dentro de los cuales se puede incluir este synset. La anotación quedaría de la siguiente forma:

```
04719796 09 n 01 Christian_Science 0 001 @ 04718274 n 0000 — religious system based on teachings of Mary Baker Eddy emphasizing spiritual healing &%ReligiousOrganization+
```

Donde el sufijo + indica que el concepto de la ontología es un hiperónimo del synset.

Instanciación. Esta última relación indica que el synset de WordNet es miembro de un concepto de la ontología. Veamos el siguiente ejemplo para la palabra *"Underground_Railroad"*:

```
00034393 04 n 02 Underground_Railroad 0 Underground_Railway 0 001 @ 00032687 n 0000 — abolitionists secret aid to escaping slaves; pre−Civil War in US
```

En este caso, el concepto más apropiado para el synset es ORGANIZATION. Por tanto, el synset asociado a *Underground_Railway*, es una organización particular. Este tipo de relación se muestra de la siguiente forma:

```
00034393 04 n 02 Underground_Railroad 0 Underground_Railway 0 001 @ 00032687 n 0000 — abolitionists secret aid to escaping slaves; pre−Civil War in US &%Organization@
```

Donde el símbolo "@" indica una relación de instanciación respecto al synset.

Mediante la inclusión de los conceptos de la ontología SUMO en WordNet, se pueden establecer relaciones entre synsets y categorías sintácticas al igual que sucedía con WordNet Domains. Además, mediante el uso de un algoritmo de desambiguación se pueden asignar a un contexto determinado los conceptos de SUMO relacionados. De esta forma, la representación conceptual

puede ser utilizada para facilitar búsquedas semánticas o clasificar documentos.

4.5 Análisis de la Semántica Latente (LSA)

El Análisis de la Semántica Latente o Latent Semantic Analysis (LSA) es un modelo computacional que explota una característica propia del lenguaje natural: palabras del mismo campo semántico tienden a aparecer juntas o en contextos similares (Landauer y Dumais (1997)).

LSA tiene sus orígenes en una técnica de recuperación de información llamada LSI (Latent Semantic Indexing) (Furnas et al. (1988), Deerwester et al. (1990)). El objetivo de LSI es mejorar la recuperación de documentos reduciendo una gran matriz de término-documento en un espacio más reducido utilizando la técnica de SVD (Singular Value Decomposition). LSA utiliza la misma metodología pero se diferencia en la representación de la matriz. En este caso, LSA utiliza una matriz palabra-contexto.

LSA representa un texto como una matriz de co-ocurrencia $M \times N$, donde las M filas se corresponden con palabras, y las N columnas se corresponden con una unidad de contexto, ya sea, una frase, un párrafo, etc. Cada celda de la matriz contiene el número de veces que una palabra determinada en la fila aparece en el contexto proporcionado por la columna. La Tabla 4.7 muestra esta representación.

Universidad de Alicante

	C1	C2	C3	C4	C5
W1	1	0	0	2	0
W2	0	4	1	0	0
W3	2	0	0	1	0

Tabla 4.7. Matriz $M_{w \times c}$

Donde:
W:palabras
C: contextos

f_{ij}: frecuencia de co-ocurrencia

LSI y LSA se diferencian principalmente en su definición de contexto utilizado. Para LSI es un documento, mientras que para LSA es más flexible, aunque a menudo hace referencia a párrafos. Si la unidad de contexto de LSA es un documento, entonces LSA y LSI son esencialmente la misma técnica.

Una vez establecida la frecuencia de co-ocurrencia de cada término con respecto a cada contexto, cada fila se interpreta como un vector contextual de d dimensiones. Donde d representa el número de contextos utilizados.

Tras insertar en las celdas correspondientes de la matriz el número de ocurrencias de cada palabra con respecto al contexto de su columna correspondiente, la matriz $M \times N$ obtenida se descompone utilizando la técnica de Singular Value Decomposition (SVD). Mediante SVD se reducen las dimensiones de la matriz inicial para que contextos similares sean redistribuidos unos dentro de otros. SVD se basa en el hecho de que cualquier matriz rectangular puede ser descompuesta en el producto de otras tres matrices. Esta descomposición puede obtenerse sin pérdida de información si no se utilizan más factores que el mínimo de N y M. En estos casos, la matriz original puede ser perfectamente reconstruida.

Sin embargo, como ocurre normalmente, LSA reduce una matriz de miles de dimensiones a unas pocas centenas. De esta forma, es prácticamente imposible reconstruir la matriz original. A pesar de que esto pueda sonar a inapropiado, es de hecho esta reducción el objetivo de LSA. El efecto de esta acción se traduce en que la pérdida de información producida es debido al ruido. De esta forma, la reducción de la dimensión produce que las relaciones de similitud entre palabras y contextos sean mucho más aparentes.

La Figura 4.7 representa de forma esquemática lo que significa la reducción de dimensiones llevada a cabo por medio de SVD. En la figura de la izquierda, cada término está representado por cuatro dimensiones, tantas como documentospárrafos existen en el corpus {d1,d2,d3,d4}. En la figura de la derecha, los términos pasan a estar representados por dos dimensiones abstractas pero de una mayor utilidad funcional. A cada término se le infiere una

probabilidad de estar representado en un concepto. En este caso, se observa cómo al término t2 se le infiere cierta probabilidad de salir en el párrafo d2 aunque como muestra la figura de la izquierda, esto no se produzca (se hace patente la característica de relacionar conceptualmente, términos con documentos aunque no aparezcan en ellos).

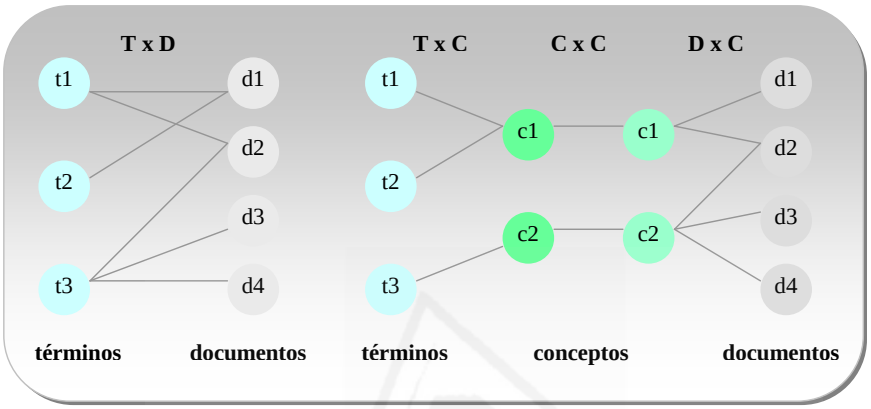


Figura 4.7. Reducción dimensional de la matriz en LSA

En otras palabras, lo que se persigue al reducir las dimensiones de la matriz original, no es más que eliminar el ruido presente en las relaciones entre términos y contextos. Esto es debido a que podemos expresar con distintos términos el mismo concepto. Además, LSA no tiene en cuenta la estructura lingüística de los contextos, simplemente las frecuencias de aparición y coocurrencias de términos.

Una analogía muy gráfica de cómo funciona la técnica, la proporciona un artículo de (Yu et al. (2004)):

"Imaginemos que tenemos un acuario de peces tropicales y tan orgullosos estamos de tenerlo que deseamos fotografiarlo para una revista especializada. Para capturar la mejor foto, elegiremos el mejor ángulo que garantice la mejor toma. Además, nos aseguraremos de que en ella salgan visibles el máximo número de peces sin ser solapados por otros peces. Tampoco queremos que los peces salgan todos juntos en una masa sino que los queremos mostrar

bien distribuidos en el agua. Como nuestro acuario es transparente, tomaremos diversas fotos desde diferentes puntos de vista y elegiremos la que mejor se adapte a lo antes descrito".

En definitiva, lo que hace la técnica es mediante la recursividad (hacer varias fotografías), buscar las dimensiones que mejor permitan una diferenciación de las "bolsas semánticas" (peces) en las que los términos participan. Una vez hecho esto, elegiremos sólo las dimensiones que mejor caractericen estas bolsas.

En la Figura 4.8 se muestra una representación de la matriz original desglosada en dos matrices de vectores singulares y una matriz diagonal de valores singulares. A partir de este desglose se reducirán las dimensiones seleccionando solamente aquellas que mejor representen las diferentes regiones semánticas. Estudios realizados han demostrado que la reducción a 300 dimensiones proporciona los mejores resultados Turney (2004).

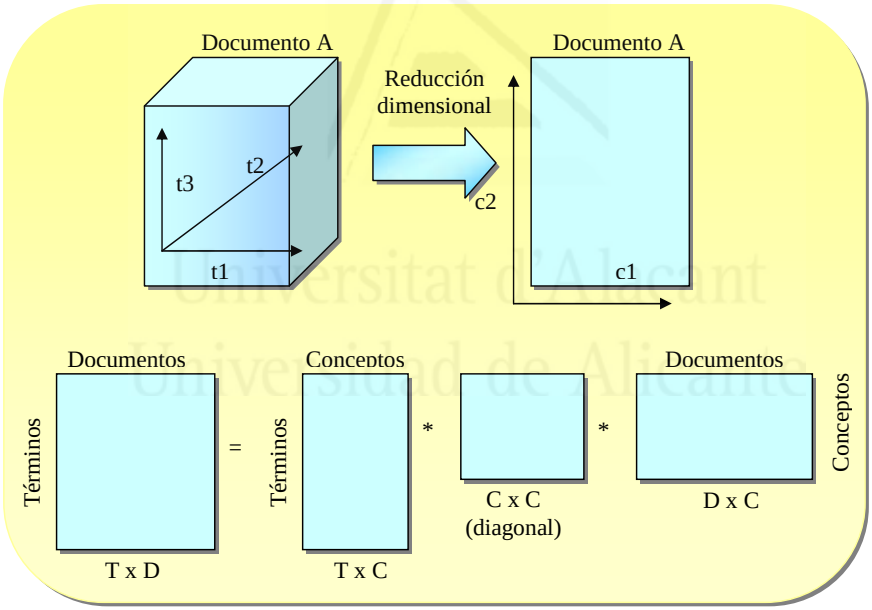


Figura 4.8. Descomposición de la matriz en LSA

En definitiva, la técnica SVD devolverá un desglose de las relaciones que se mantienen en la matriz original. De esta forma,

podremos reconstruir la matriz inicial pero tomando en consideración sólo las dimensiones que hacen más fuerte la relación entre términos y documentos. Esto se hará tomando los valores singulares más altos y volviendo a multiplicar las tres matrices pero reduciendo sus dimensiones a las mismas que valores singulares hayamos considerado.

En (Landauer y Dumais (1997)) se evaluó la habilidad de LSA para reconocer sinónimos. En este estudio, se utilizó el Test of English as a Foreign Language (TOEFL) como base para la evaluación del funcionamiento de LSA. El TOEFL es una prueba que consiste en dada una palabra clave, seleccionar de entre cuatro opciones (palabras) distintas, cuál es más similar a la palabra clave. Por ejemplo, si *"distant"* es la palabra clave en cuestión y las cuatro opciones son: 1. *"impossible"*, 2. *"faraway"*, 3. *"observable"*, 4. *"fearful"*. En este caso, se debería seleccionar la opción 2 como la más similar en cuanto a significado.

Para poder aplicar LSA sobre el tipo de información proporcionada por TOEFL y construir la matriz contextual, se utilizó como corpus la versión electrónica de la Grolier's Academic American Encyclopedia. En esta enciclopedia habían alrededor de 30473 artículos, para cada uno de los cuales se extrajeron los primeros 2000 caracteres, con una media de 151 palabras por artículo. Una vez obtenida esta información, los datos se introdujeron en la matriz contextual, teniendo ésta una columna por artículo (30473 columnas) y alrededor de 60768 filas, donde cada fila se correspondía con una palabra que aparecía en al menos dos artículos (columnas) de la matriz. Inicialmente, las celdas de la matriz contenían la frecuencia en que una palabra aparecía en cada artículo. Posteriormente, estos datos se transformaron según la fórmula $\ln(1+frec)/$(entropía de la palabra sobre todos los contextos). La matriz resultante final fue reducida hasta 300 dimensiones mediante la técnica de SVD, correspondiéndose estas dimensiones con los valores singulares más elevados obtenidos, dando como resultado vectores con 300 valores reales para representar cada palabra. La similitud entre dos palabras se midió utilizando el coseno entre vectores, de forma que cuanto más similares fueran dos palabras mayor sería el coseno obtenido entre ambas. Por lo

tanto, para seleccionar la respuesta correcta de cada pregunta se calculó el coseno entre la palabra y cada una de las cuatro alternativas. Como resultado, la aplicación de LSA sobre el test de TOEFL supuso un acierto del 65 %, puntuación similar a los resultados obtenidos por personas de habla no inglesa que realizaron el test. Esta misma prueba se realizó sobre la matriz original de 30000×60000, en este caso, la precisión bajó a un 37 %, sugiriendo estos resultados que la descomposición de la matriz eliminaba el ruido presente en la matriz original, proporcionando una mejor representación de las similitudes entre palabras.

El test TOEFL fue utilizado de nuevo por (Turney (2001)) para evaluar los resultados de un sistema que empleaba una técnica distinta a LSA. En este caso, el sistema empleado calculaba los valores de la Pointwise Mutual Information (PMI) entre la palabra dada y sus cuatro posibilidades, utilizando los resultados de la búsqueda en Alta Vista como base para establecer la frecuencia de aparición de las palabras. La estrategia utilizada para calcular el grado de similitud entre la palabra clave y las cuatro opciones se muestra en la Tabla 4.8. En este ejemplo se trata de averiguar la palabra más similar a *"levied"* de entre estas 4 opciones: *"imposed"*, *"believed"*, *"requested"*, *"correlated"*. Según los resultados obtenidos la palabra con mayor similitud semántica es *"imposed"*.

Este mismo ejemplo aplicando la técnica de LSA obtiene el mismo resultado, tal y como muestra la Tabla 4.9.

Es interesante destacar que a pesar de utilizar diferentes fuentes de información ambas aproximaciones obtienen el mismo resultado. La comparativa sobre las 80 preguntas del TOEFL para cada aproximación se muestra en la Tabla 4.10.

Como demuestran los resultados, PMI mejora en un 10 % los resultados de LSA. Pero la interpretación de estos resultados es complicada debido a que ambas técnicas utilizan fuentes de información muy distintas y PMI utiliza un contexto mucho más pequeño que LSA.

Estos ejemplos, sirven para mostrar dos aproximaciones basadas en la utilización de grandes cantidades de información para establecer relaciones semánticas obteniendo resultados muy posi-

Búsqueda	Resultados
imposed AND NOT (imposed NEAR "not")	1,147,535
believed AND NOT (believed NEAR "not")	2,246,982
requested AND NOT (requested NEAR "not")	7,457,552
correlated AND NOT (correlated NEAR "not")	296,631
(levied NEAR imposed) AND NOT ((levied OR imposed) NEAR "not")	2,299
(levied NEAR believed) AND NOT ((levied OR believed) NEAR "not")	80
(levied NEAR requested) AND NOT ((levied OR requested) NEAR "not")	216
(levied NEAR correlated) AND NOT ((levied OR requested) NEAR "not")	3

Selección	Similitud
$p(levied\|imposed)$	0.0020034
$p(levied\|believed)$	0.0000356
$p(levied\|requested)$	0.0000290
$p(levied\|correlated)$	0.0000101

Tabla 4.8. Cálculo similitud PMI para TOEFL

Selección	Resultado LSA
imposed	0.70
believed	0.09
requested	0.05
correlated	-0.03

Tabla 4.9. Resultado LSA sobre TOEFL

$P(pal\_clave\|opcion)$	Respuestas correctas	Porcentaje
PMI	59/80	73.75 %
LSA	51.5/80	64.4 %
Persona de habla no inglesa	51.6/80	64.5 %

Tabla 4.10. Comparativa LSA y PMI sobre TOEFL

tivos. En el siguiente capítulo estas técnicas han sido adaptadas y utilizadas en distintos métodos de desambiguación automática.

Universitat d'Alacant
Universidad de Alicante

Capítulo 5

Métodos

En este capítulo se describen los métodos desarrollados a partir de los estudios realizados en esta tesis: WSD DRelevant, WSD DLSA y WSD SenseDiscrim. Todos los métodos descritos a continuación se clasifican dentro del grupo de métodos basados en conocimiento, ya que, la información necesaria para su correcto funcionamiento procede de corpus y recursos léxicos y no requieren ningún proceso de aprendizaje. Además de la descripción de cada método también se presentan las diferentes aproximaciones realizadas, así como la utilización de los diferentes recursos descritos en el capítulo anterior.

Universitat d'Alacant
Universidad de Alicante

5.1 WSD basado en conocimiento: DRelevant

El método desarrollado en esta sección se clasifica dentro de los métodos no supervisados basados en conocimiento. Dentro de esta categoría encontramos aquellos métodos que necesitan de información externa procedente de diversas fuentes, ya sea de diccionarios electrónicos, corpus generales, corpus especializados, etc. A partir de esta información los métodos no supervisados pueden obtener los datos necesarios para construir sus propias fuentes de conocimiento y relacionar de esta forma palabras a partir de sus contextos y sus apariciones junto a otras palabras semánticamente relacionadas.

La idea principal en la que se basa la implementación del método descrito a continuación, es en la utilización de una serie de categorías semánticas asociadas a los sentidos de las palabras como una aproximación para determinar los sentidos de éstas. Es decir, a partir de WordNet Domains que está etiquetado con una serie de categorías o dominios, se extraen de forma estadística los contextos en los que aparecen las distintas palabras a desambiguar, y se determina a qué categoría semántica pertenecen esos contextos.

La base teórica en la que se basa este sistema parte de tres premisas:

- **Primera:** Las palabras que pertenecen a diferentes clases conceptuales o dominios, como ANIMAL o MÁQUINA, tienden a aparecer en contextos bien diferenciados.
- **Segunda:** Los diferentes sentidos de una palabra tienden a pertenecer a clases conceptuales diferentes.
- **Tercera:** Si se puede construir un discriminador de contextos para diferentes clases conceptuales entonces podremos distinguir los sentidos de las palabras que pertenecen a esas clases.

En la Tabla 5.1 podemos apreciar la diferencia existente entre los contextos donde aparece la palabra *"crane"* con el sentido asociado de "grúa" y *"crane"* con el sentido de "grulla". Ambas acepciones de la misma palabra pertenecen a categorías semánticas distintas. La primera acepción pertenece a la categoría INDUSTRY y la segunda a ZOOLOGY.

Universitat d'Alacant
Universidad de Alicante

Contexto de entrada	Categoría semántica
Treadmills attached to "cranes" were used to lift heavy — for supplying power for "cranes", hoists, and lifts — above this height, a tower "crane" is often used.	INDUSTRY
This elaborate courtship rituals "cranes" build a nest of vegetation — are more closely related to "cranes" and rails. They ran — low trees. At least five "crane" species are in danger.	ZOOLOGY

Tabla 5.1. Contextos asociados a diferentes sentidos de la palabra "crane"

En este caso, a partir del contexto que rodea a la palabra *"crane"* podemos seleccionar su sentido correspondiente, utilizando para ello la información procedente de las palabras que la rodean. La utilización de las palabras del contexto como fuente de información para establecer el sentido de una palabra y las relaciones semánticas entre las palabras del contexto, son la base de nuestro método de desambiguación automática.

5.1.1 Obtención y categorización de contextos

El objetivo principal en este punto es poder establecer relaciones entre palabras a partir de los contextos en los cuales aparecen. Si esos contextos se engloban dentro de una categoría determinada, podemos establecer, analizando la frecuencia de aparición de las palabras, si éstas pertenecen o no a esa categoría. Para ello, es necesario recopilar cierta información acerca de su frecuencia de aparición junto a otras palabras y las categorías en las que suelen aparecer.

La estrategia seguida para recopilar información relativa a la frecuencia de aparición consta de tres pasos:

1. Extraer los contextos representativos para cada categoría.
2. Identificar las palabras más destacadas entre los contextos obtenidos y establecer un determinado peso para cada palabra.
3. Utilizar los pesos resultantes para poder determinar la categoría apropiada de una palabra polisémica (con más de un sentido) dentro de un nuevo contexto.

La existencia de palabras polisémicas supone un problema para el establecimiento de los contextos apropiados para cada categoría. Podemos encontrar por ejemplo, instancias de una misma palabra que corresponden a sentidos diferentes. En este caso, podemos adquirir información errónea y que produce ruido debido a estas palabras. Normalmente, se puede amortiguar el efecto de estos errores de clasificación porque suelen distribuirse entre varias categorías. Sin embargo, si una misma palabra atiende a varios sentidos y éstos se distribuyen dentro de la misma categoría, daría lugar a una clasificación errónea. Además, puede ser que uno de

los sentidos de la palabra polisémica aparezca con mayor frecuencia en los contextos extraídos, siendo este sentido incorrecto a la categoría que estamos contextualizando.

Para minimizar los efectos que producen la aparición de palabras muy frecuentes dentro de cada categoría, se realiza una ponderación de los contextos. Es decir, si una palabra aparece muy frecuentemente en una categoría no por ello es significativa de esa categoría, ya que, esa misma palabra puede aparecer muy frecuentemente junto a otras categorías. Para evitar este tipo de situaciones se utiliza la siguiente ponderación: si una palabra aparece k-veces en el corpus, entonces las palabras que la rodean contribuyen $1/k$ a la suma de la frecuencia.

5.1.2 Extracción de contextos

Para poder realizar estimaciones estadísticas de co-ocurrencia de palabras en diferentes contextos, es necesario disponer de un corpus previamente categorizado. El problema que encontramos actualmente es que no existen corpus de ámbito general lo suficientemente extensos para satisfacer la demanda de relaciones entre palabras y categorías. Además, la mayoría de corpus existentes se centran en dominios de ámbito de aplicación específicos. Debido a estas dificultades, es necesaria la utilización de algún otro recurso que recoja de forma genérica las distintas relaciones entre palabras y su pertenencia a diferentes categorías semánticas.

Observando las prestaciones de distintos recursos electrónicos como WordNet Domains o SUMO, donde a partir de las definiciones de un diccionario electrónico se añaden etiquetas semánticas a las palabras, se ha optado por utilizar estos recursos como corpus para la categorización de contextos. La estructura de estos recursos ya fue presentada en el capítulo anterior y a modo de recordatorio se muestra un pequeño extracto de su configuración en las siguientes figuras.

La Figura 5.1 muestra cómo con WordNet Domains a partir de las distintas categorías semánticas Economy, Industry y Botany, se pueden clasificar tanto palabras pertenecientes a

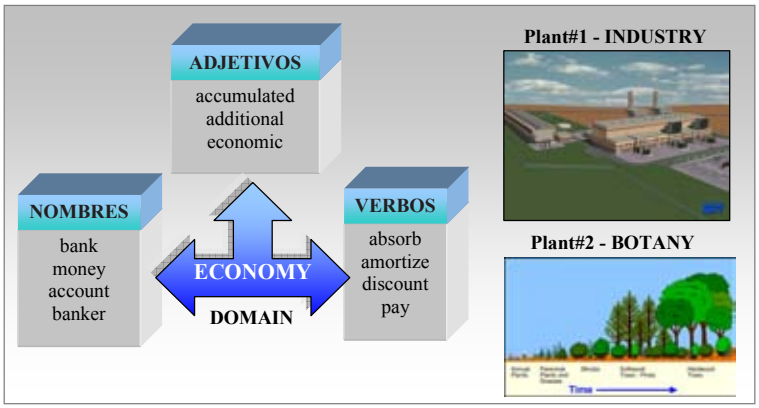


Figura 5.1. WordNet Domains

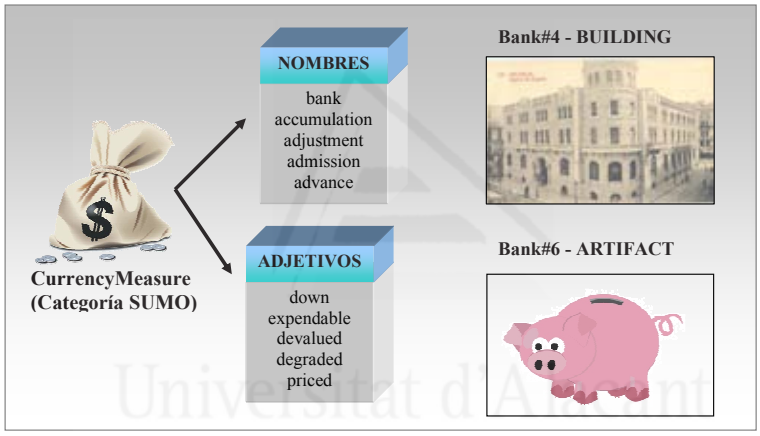


Figura 5.2. SUMO

Universitat d'Alacant
Universidad de Alicante

diferentes categorías sintácticas, como determinar los diferentes sentidos de una palabra polisémica.

En la Figura 5.2 se muestra cómo en la ontología SUMO también se pueden agrupar palabras de distintas categorías sintácticas con la misma etiqueta semántica (CurrencyMeasure). Salvo que en este caso, no se encuentran verbos relacionados con CurrencyMeasure, esto es debido a que esta ontología presenta un grado de granularidad más fino que el caso de WordNet Domains y para verbos que en WordNet domains compartían la

etiqueta ECONOMY ahora en SUMO tienen otras categorías asignadas: FINANCIALTRANSACTION: *"pay", "amortize",* GETTING *"absorb"* y DECREASING *"discount"*.

A pesar de sus diferencias en cuanto a anotación, mediante ambos recursos se puede extraer información relacionada con los contextos de cada categoría, utilizando para ello las glosas de WordNet (ya que ambos se basan en su base de datos léxica). Como ya se comentó en el capítulo anterior, en WordNet, cada palabra tiene asociada una definición (glosa) que puede además incluir algunos ejemplos de uso de la palabra que se está definiendo. Por ejemplo, el caso de la palabra *"bank"* con la categoría ECONOMY su glosa es: *"the funds held by a gambling house or the dealer in some gambling games; "he tried to break the* ***bank*** *at Monte Carlo" "*. En este caso, además de dar una definición para uno de los sentidos de *"bank"* también se proporciona un ejemplo de una frase donde se utiliza *"bank"* con ese sentido.

Para poder determinar los contextos de cada categoría se han agrupado todas las glosas que tienen asociada una misma etiqueta semántica. De esta forma, se ha obtenido una clasificación por categorías de todas las glosas de WordNet, tanto usando las categorías de WordNet Domains como usando las categorías de SUMO.

5.1.3 Obtención de las palabras significativas

Una vez determinados los contextos asociados a cada categoría, es necesario determinar de entre todas las palabras que aparecen en las glosas, cuáles son las más significativas en relación a esa categoría.

Cuando hablamos de palabras significativas nos referimos a aquellas palabras que aparecen más a menudo en el contexto de una categoría que en cualquier otra parte del corpus. En este caso, se podría determinar la importancia de una palabra con respecto a una categoría atendiendo a la Fórmula 5.1:

$$importancia(w) = \frac{P(palabra|categoria)}{P(palabra)} \tag{5.1}$$

WordNet Domains
Etiqueta:ECONOMY
01517979—absorb— take up, as of debts or payments; "absorb the costs for something"
01549722—account— keep an account of
00106728—accrue— grow by addition, as of capital: "The interest accrues"
01560796—advance— pay in advance; "Can you advance me some money?"
00108725—advance— rise in rate or price; "The stock market gained 24 points today"
00810882—afford— be able to spare or give up; "I can't afford to spend two hours with this person"
00490924—allow— give or assign a share of money or time to a particular person or cause; "I will earmark this money...
01585121—allow— grant as a discount or in exchange; "The camera store owner allowed me $50 on my old camera"
01606528—amortise— liquidate gradually
01606528—amortize— liquidate gradually
...

Figura 5.3. Clasificación contextual a partir de WordNet Domains

SUMO
Etiqueta:FINANCIALTRANSACTION
01544212—cash—exchange for cash
01544212—cash_in—exchange for cash
01544337—redeem—convert into cash; of commercial papers
01544440—redeem—pay off, as of loans or promissory notes
01544440—pay_off—pay off, as of loans or promissory notes
01544554—ransom—exchange or buy back for money; under threat
01544554—redeem—exchange or buy back for money; under threat
01546904—arbitrage—practice arbitrage, as of stocks
01547100—turn_over—do business worth a certain amount of money
01547218—broker—act as a broker
...

Figura 5.4. Clasificación contextual a partir de SUMO

La probabilidad de aparición de una palabra sobre una categoría $P(palabra|categoria)$ normalmente se suele estimar contando el número de ocurrencias de la palabra en los distintos contextos de la categoría (contexto local). Pero esta medida no da resultados reales cuando por ejemplo, una palabra es poco frecuente. Por lo tanto, se puede suavizar el efecto añadiendo la frecuencia global de aparición de la palabra $P(palabra)$ (contexto global) (Yarowsky (1992)). De esta forma, se evita obtener estimaciones erróneas a partir del contexto local, ya que, los errores que puedan aparecer dentro del contexto global son irrelevantes.

En la Tabla 5.2 se muestra el resultado obtenido tras analizar la frecuencia de aparición de la palabra *"plant"* como nombre en WordNet Domains.

5.1.4 Similitud semántica

Una vez estimada la importancia de cada palabra sobre cada categoría del corpus, es necesario establecer una medida que determine el grado de similitud semántica entre cualquier par de palabras, es decir, si existe algún tipo de relación que las vincule semánticamente.

La similitud entre palabras se define normalmente en términos de co-ocurrencia estadística. Se debe matizar que la co-ocurrencia se puede especificar de diferentes formas. De entre todas las formas posibles tenemos dos tipos principales: la co-ocurrencia medida a partir de las relaciones entre palabras dentro de un rango específico (ventana) y la co-ocurrencia de palabras como un patrón particular de relaciones gramaticales tales como sujeto-verbo y verbo-objeto.

Podemos referirnos a los dos tipos de co-ocurrencia anteriores como medidas de co-ocurrencia de primer orden. Esta denominación nos servirá para distinguirlas de aquellas medidas a las que denominaremos co-ocurrencia de segundo orden (o indirecta). Este último tipo representa una co-ocurrencia de pares de palabras en expresiones comunes. La hipótesis asociada a la co-ocurrencia de primer orden es que las palabras relacionadas semánticamente tienden a aparecer en contextos de ámbito restringido. Mientras que la idea subyacente para la co-ocurrencia de segundo orden es que las palabras semánticamente similares tienen tendencia a compartir contextos similares. Por ejemplo, "cerveza" y "vino" se consideran palabras similares porque están relacionadas semánticamente con el verbo "beber", ya que, frecuentemente aparecen como su objeto directo y por tanto, comparten un contexto similar.

La co-ocurrencia de un par de palabras (o expresiones) puede medirse de muchas formas. La medida más simple es la *"bare co-occurrence frequency"*. Otras medidas de co-ocurrencia más

Dominio	Frecuencia	Importancia
administration	4	0,001025904
agriculture	77	0,019748654
alimentation	37	0,009489613
anatomy	42	0,010771993
anthropology	8	0,002051808
archaeology	6	0,001538856
architecture	1	0,000256476
art	3	0,000769428
biology	735	0,188509874
body_care	3	0,000769428
botany	2306	0,591433701
building_industry	19	0,004873044
chemistry	85	0,021800462
color	13	0,003334188
commerce	3	0,000769428
ecology	2	0,000512952
economy	3	0,000769428
electricity	1	0,000256476
engineering	1	0,000256476
enterprise	15	0,00384714
entomology	82	0,021031034
factotum	112	0,028725314
fashion	1	0,000256476
gastronomy	84	0,021543986
genetics	1	0,000256476
geography	7	0,001795332
geology	4	0,001025904
industry	39	0,010002565
medicine	49	0,012567325
meteorology	1	0,000256476
military	1	0,000256476
mountaineering	1	0,000256476
pharmacy	30	0,007694281
physics	1	0,000256476
physiology	6	0,001538856
psychology	4	0,001025904
pure_science	1	0,000256476
quality	7	0,001795332
religion	1	0,000256476
sexuality	1	0,000256476
telecommunication	2	0,000512952
theatre	1	0,000256476
time_period	2	0,000512952
town_planning	1	0,000256476
transport	2	0,000512952
zoology	83	0,02128751
zootechnics	11	0,002821236

Tabla 5.2. Frecuencia de *"plant"* en WordNet Domains

Universitat d'Alacant
Universidad de Alicante

sofisticadas para pares de palabras están basadas en sus frecuencias de co-ocurrencia y frecuencias independientes. Una medida muy conocida de co-ocurrencia que fue inicialmente utilizada por (Church y Hanks (1990)) es la llamada Información Mutua (IM) definida según la ecuación 5.2:

$$IM(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} = \log_2 \frac{\frac{freq(x,y)}{N}}{\frac{freq(x)}{N}\frac{freq(y)}{N}} = \log_2 \frac{freq(x,y)N}{freq(x)freq(y)} \tag{5.2}$$

En la fórmula de la IM $P(x,y)$ y $freq(x,y)$ son la probabilidad de co-ocurrencia y la frecuencia de dos eventos x e y, $P(x)$ y $freq(x)$ son la probabilidad independiente y frecuencia de x, y N es el número total de eventos.

Dentro de la lexicografía computacional podemos utilizar el valor proporcionado por la IM para establecer si existe algún tipo de relación entre diferentes palabras. De forma que:

Si $IM(w_1, w_2) \gg 0$. Significa que la palabra w_1, aparece junto a la palabra w_2 más a menudo de lo que a simple vista podría parecer.

Si $IM(w_1, w_2) \ll 0$. Significa que la palabra w_1, aparece junto a la palabra w_2 menos frecuentemente de lo que cabría pensar.

Si $IM(w_1, w_2) \approx 0$. Indica que no existe ninguna evidencia que relacione la palabra w_1 con la palabra w_2.

Universidad de Alicante

Por ejemplo, en la tabla 5.3 se muestra la Información Mutua de la palabra *"drink"* junto con sus posibles objetos directos extraídos de una serie de documentos:

Podemos observar que existen palabras que aparecen con mucha frecuencia junto a *"drink"* como son: *"water"*, *"beer"*, *"alcohol"*... Pero sin embargo, el valor dado por la IM es menor que el obtenido para otras palabras cuya frecuencia de aparición es menor como: *"martinis"*, *"cup of water"*, *"champagne"*... Esta peculiaridad es debida a que la IM mide las relaciones semánticas entre palabras no sólo a partir de su frecuencia de aparición local, sino también de su frecuencia de aparición global. Debido a

Objeto directo	IM	Frecuencia
martinis	12,6	3
cup of water	11,6	3
champagne	10,9	3
beverage	10,8	8
cup of tea	10,6	2
cognac	10,6	2
beer	9,9	29
cup	9,7	6
coffee	9,7	12
toast	9,6	4
alcohol	9,4	20
wine	9,3	10
fluid	9,0	5
liquor	8,9	4
tea	8,9	5
milk	8,7	8
juice	8,3	4
water	7,2	43
quantity	7,1	4

Tabla 5.3. IM de la palabra "drink" con sus posibles objetos directos

ello, las palabras que aparecen casi exclusivamente junto a otras tenderán a tener una relación semántica más fuerte que otras palabras demasiado frecuentes en el corpus, lo que indica que tienen un uso genérico y por tanto, una relación semántica más débil.

La medida de la Información Mutua ha sido utilizada para medir la similitud entre palabras (Dagan et al. (1993), (Lin (1998a)), (Pekar y Krkoska (2003)), extracción de sinónimos (Turney (2001)) y también para extracción de colocaciones (Church et al. (1991)).

En este trabajo, se utiliza la fórmula de la Información Mutua en una forma más relajada Yarowsky (1992), es decir, nos interesan aquellas palabras que aparecen más a menudo en el contexto de un dominio, con respecto a otros (palabras significativas con respecto a un dominio).

$$IM(w, D) = \log_2 \frac{P(w, D)}{P(w)P(D)} = \log_2 \frac{P(D|w)}{P(D)} = \log_2 \frac{P(w|D)}{P(w)} \tag{5.3}$$

En la Tabla 5.4 se muestran los resultados obtenidos tras el cálculo de la Información Mutua de la palabra "plant" en WordNet Domains.

5.1.4.1 Perfeccionamiento de la Información Mutua.

El valor obtenido por la fórmula de la Información Mutua 5.3, mide la importancia de las palabras frente a un dominio de WND. Pero este valor no refleja exactamente la relevancia que tiene una palabra con respecto a un determinado dominio. Es por tanto necesario establecer la frecuencia local de una palabra con respecto a cada dominio para poder establecer de forma específica la relevancia de esa palabra.

Esta nueva forma de establecer la relevancia de una palabra w sobre un dominio D se mide a través de la fórmula del Ratio de Asociación (RA) (Rigau (1998), Framis (1994)).

La fórmula del Ratio de Asociación se muestra en 5.4.

$$RA(w, D) = Pr(w|D) \log_2 \frac{Pr(w|D)}{Pr(w)} \tag{5.4}$$

La Tabla 5.5, muestra el resultado tras la unión de los dos conceptos: importancia (IM) y relevancia (RA) de una palabra frente a un dominio. Los resultados demuestran que existen casos para los que la IM asigna mayor peso y que ahora mediante el RA quedan relegados a posiciones inferiores. Véanse por ejemplo, los resultados referidos al dominio Religion en ambas Tablas 5.4 y 5.5:

Religion−5,904181(IM)
Religion−0,000472(RA).

Queda patente por tanto, la necesidad de identificar correctamente la relevancia de una palabra con respecto a un dominio, para así evitar una ponderación errónea usando la medida IM, que no tiene en cuenta la frecuencia local asociada ($P(w|D)$).

Dominio	IM
religion	5,904181
military	5,799872
physics	5,30723
fashion	4,950001
administration	4,596772
transport	4,254565
geography	4,220493
town planning	4,076359
time period	3,940711
economy	3,877388
quality	3,611403
electricity	3,471225
psychology	3,451331
sexuality	3,444749
meteorology	3,292426
geology	3,08912
factotum	3,076359
agriculture	2,8921
commerce	2,890722
theatre	2,793567
zootechnics	2,701343
art	2,661958
botany	2,552171
biology	2,449791
architecture	2,345
telecommunication	2,172187
physiology	1,908697
zoology	1,611403
entomology	1,589572
building industry	1,575955
archaeology	1,473461
mountaineering	1,457541
alimentation	1,071378
anatomy	0,978776
medicine	0,838633
ecology	0,768773
body care	0,701079
engineering	0,65199
anthropology	0,545312
pure science	0,508863
industry	0,459295
pharmacy	0,389059
enterprise	0,323433
color	0,320542
genetics	0,277378
gastronomy	0,211381
chemistry	0,017062

Tabla 5.4. IM para *"plant"* en WND

Dominio	RA
agriculture	0,102860
botany	0,071716
biology	0,064123
entomology	0,022920
archaeology	0,019603
mountaineering	0,019178
alimentation	0,010787
ecology	0,006275
industry	0,003025
building_industry	0,002533
pharmacy	0,002441
physiology	0,002435
anatomy	0,002379
medicine	0,002247
architecture	0,002211
body_care	0,002066
art	0,002015
engineering	0,001988
enterprise	0,001939
color	0,001918
commerce	0,001867
anthropology	0,001790
factotum	0,001747
geology	0,001739
meteorology	0,001610
electricity	0,001500
economy	0,001264
gastronomy	0,001173
genetics	0,001096
geography	0,001085
administration	0,000910
fashion	0,000767
physics	0,000642
military	0,000499
chemistry	0,000083
psychology	0,001512
zootechnics	0,084177
zoology	0,002527
telecommunication	0,002309
theatre	0,001930
pure_science	0,001713
sexuality	0,001516
psychology	0,001512
quality	0,001416
town_planning	0,001158
transport	0,001068
religion	0,000472

Tabla 5.5. RA para *"plant"* en WND

Universitat d'Alacant
Universidad de Alicante

5.1.5 Vectores de co-ocurrencia

Una vez establecida la medida de co-ocurrencia para determinar la similitud entre palabras y dominios ($RA(w|D)$), es necesario hallar una forma de obtener una medida de similitud entre pares de palabras, contextos, etc. Para este tipo de tarea se utilizan los vectores de co-ocurrencia.

Veamos con un ejemplo sencillo la forma de obtener vectores de co-ocurrencia para una palabra determinada. Supongamos que encontramos el texto de la Figura 5.5.

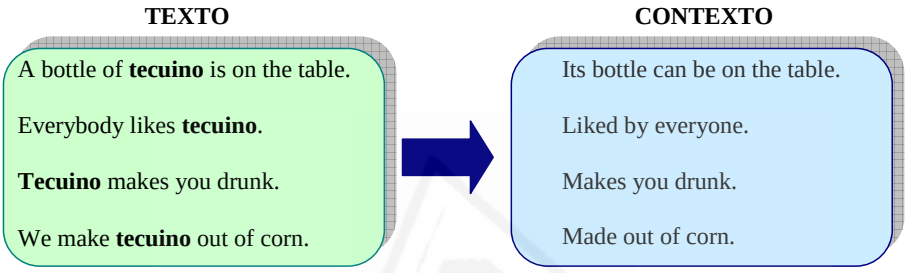


Figura 5.5. Determinación del significado de *"tecuino"*

Dada la palabra *"tecuino"* de la cual no sabemos su significado, se puede inferir su sentido de acuerdo a las palabras que la rodean: "You shall know a word by the company it keeps!" (Firth (1957)).

En este caso, a partir del contexto podemos deducir que el *"tecuino"* podría ser una bebida alcohólica hecha de maíz. Esta idea intuitiva puede capturarse utilizando un vector cuyas componentes podrían ser las palabras que rodean a la palabra *"tecuino"*, para posteriormente compararlo con vectores de otras palabras como por ejemplo "beer", "wine" o "tequila". De esta forma, se podría determinar que cualquier tipo de bebida alcohólica comparte muchas de las componentes de la palabra en cuestión, concluyendo por tanto: *"tecuino"* $\equiv$ bebida alcohólica.

A partir de esta idea intuitiva se puede representar cualquier palabra con un vector de características (features). Estas características pueden ser las palabras del contexto (usando una ventana de un tamaño específico), las categorías sintácticas, los

objetos directos, etc. Por ejemplo, supongamos que tenemos f_i características binarias que indican si una palabra w está (1) o no (0) en un contexto X de N palabras. Podríamos representar mediante un vector cualquier palabra w, de forma que: $\overrightarrow{w} = \{f_1, f_2, f_3, ...f_N\}$. Si la palabra w = *"tecuino"* y las palabras del contexto fueran *"bottle"*, *"drunk"* y *"matrix"*, el vector obtenido sería: $\overrightarrow{w} = \{1, 1, 0, ...\}$.

Entonces, dadas cualquier par de palabras (w_1, w_2), cada una representada por su vector de características $(\overrightarrow{w_1}, \overrightarrow{w_2})$, se podría determinar su similitud mediante la utilización de una medida que calcule la proximidad de los dos vectores de características.

En la Tabla 5.6 se muestra de forma intuitiva los vectores de co-ocurrencia de cuatro palabras *"apricot" "pineapple"*, *"digital"* e *"information"*. Estos ejemplos han sido extraídos utilizando un contexto de dos líneas del Brown Corpus.

Partiendo de esta representación es evidente la necesidad de una medida que establezca un grado de similitud muy elevado entre *"apricot"* y *"pineapple"* y entre *"digital"* e *"information"*, y además devuelva un valor de similitud muy bajo entre los pares de palabras que no tienen nada en común.

	arts	boil	data	function	large	sugar	summarized
apricot	0	1	0	0	1	1	0
pineapple	0	1	0	0	1	1	0
digital	0	0	1	1	1	0	1
information	0	0	1	1	1	0	1

Tabla 5.6. Vectores de co-ocurrencia Brown Corpus

Como conclusión se extrae que para determinar el grado de similitud entre dos palabras es necesario:

1. Definir el tipo de características utilizadas para crear los vectores.
2. Definir la ventana contextual de extracción de información.
3. Determinar el valor que se le da a cada característica del vector (binario, frecuencia, Información Mutua...)

4. Métrica que establezca la distancia entre dos vectores (coseno, distancia euclídea...)

5.1.5.1 WND y SUMO como características.

En el ejemplo anterior se han utilizado como características para construir el vector de co-ocurrencia, las palabras que aparecen en un contexto de tamaño N. La aproximación que se propone en este trabajo es la utilización de las categorías semánticas de WND y SUMO como características para construir los vectores de co-ocurrencia. De esta forma, se evita la creación de vectores con un elevado número de características, muchas veces irrelevantes a la hora de establecer la similitud entre dos palabras.

A continuación se muestra un ejemplo para WND:

Dada la frase:

"There are a number of ways in which the chromosome structure can change, which will detrimentally change the genotype and phenotype of the organism"

Los pasos a seguir para obtener el vector de características son:

- **Extracción de palabras del contexto.** Se extraen aquellas palabras con contenido semántico, omitiendo las stop words (artículos, preposiciones, ...). El resultado es una lista con: {*number, way, chromosome structure change detrimentally genotype phenotype organism*}.
- **Asignación de dominios relevantes.** Para cada una de estas palabras se extraen sus dominios relevantes (calculados previamente con el Ratio de Asociación).
- **Ponderación de dominios.** Se establecen cuáles son los dominios más relevantes del contexto. Para ello, se ponderan aquellos dominios que aparecen en mayor número de palabras tal y como se muestra en la Fórmula 5.5.

$$Peso\_Dom_i = \sum_{Dom\_w_i}^{Dom\_w_N} RA_{Dom_i} \tag{5.5}$$

Los pesos de aquellos dominios que aparecen en distintas palabras se van agregando de forma que, finalmente se obtiene una

lista de dominios no repetidos, de las palabras que intervienen en el contexto.

La Figura 5.6 muestra el vector de características obtenido para el ejemplo anterior.

Biology 0.03102837
Ecology 0.00402855
Botany 0.00003204
Zoology 0.00001779
Anatomy 0.00001295
Physiology 0.000001002
Chemistry 0.00000100017
...

Figura 5.6. Vector de co-ocurrencia usando WND

Una vez generado el vector de características, es necesario medir la similitud entre diferentes vectores utilizando alguna métrica específica de comparación de vectores.

5.1.6 Métricas sobre vectores

Gracias al vector de características cuya obtención se ha comentado en la sección anterior, es posible determinar el grado de similitud entre dos palabras o contextos.

Para definir el grado de similitud entre dos palabras w y v, se necesita una métrica que tenga en cuenta los vectores asociados a $\overrightarrow{w}$ y $\overrightarrow{v}$ y obtenga un valor de similitud entre ambos. Las métricas más simples utilizadas para medir la distancia (o similitud) entre vectores son la Manhattan y la distancia Euclídea. En la Figura 5.7 se muestra un gráfico intuitivo de ambas métricas aplicadas sobre dos vectores bidimensionales.

La distancia Manhattan también conocida como la distancia de Levenshtein o norma L1 es:

$$Manhattan(\overrightarrow{x}, \overrightarrow{y}) = \sum_{i=1}^{N} |x_i - y_i| \tag{5.6}$$

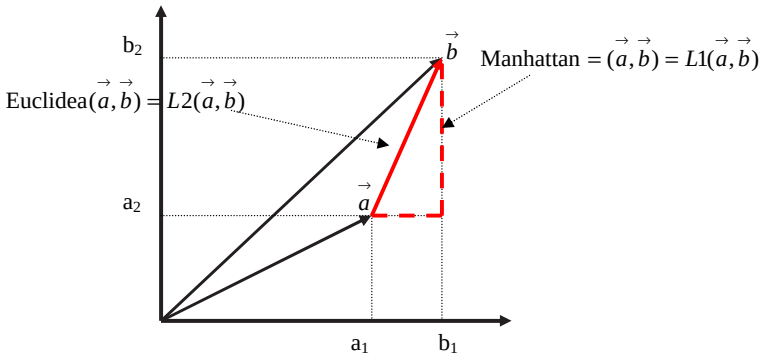


Figura 5.7. Distancia Euclídea y Manhattan

La distancia Euclídea se define como:

$$Euclidea(\overrightarrow{x}, \overrightarrow{y}) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \tag{5.7}$$

A pesar de que estas dos métricas proporcionan una medida de similitud muy intuitiva entre vectores, han sido muy poco utilizadas para medir similitud entre palabras. Esto es debido a que ambas métricas son muy sensibles a valores extremos.

En lugar de utilizar estas métricas tan sencillas, la similitud entre palabras está estrechamente relacionada con las métricas utilizadas en Extracción de Información. Dado que los sistemas de extracción de información funcionan bastante bien a la hora de establecer la similitud entre palabras vamos a describir algunas de estas métricas.

Utilizando como ejemplo el vector de características binarias mostrado en la Tabla 5.6, donde la similitud entre dos vectores era justamente el número de características que tenían en común, podemos asumir que tenemos un vector binario. De esta forma, se define la métrica de similitud utilizando el producto escalar, como sigue:

$$Prod\_escalar = (\overrightarrow{v}, \overrightarrow{w}) = \overrightarrow{v} \cdot \overrightarrow{w} = \sum_{i=1}^{N} v_i \times w_i \tag{5.8}$$

Esta nueva métrica asume que los vectores son binarios, pero ocurre como hemos visto en la sección previa, que las características pueden almacenar valores de asociación entre palabras no binarios. Para este caso genérico el vector definido para una palabra $\overrightarrow{w}$ con N características $f_1 ... f_N$ es como sigue:

$$\overrightarrow{w} = (asoc(w, f_1), asoc(w, f_2), ...asoc(w, f_N)) \tag{5.9}$$

Ahora se puede aplicar la métrica del producto escalar sobre vectores que contienen valores de asociación, en lugar de valores binarios. Pero el resultado proporcionado tiene un problema: favorece a los vectores más largos. La longitud de un vector se define como:

$$|\overrightarrow{v}| = \sqrt{\sum_{i=1}^{N} v_i^2} \tag{5.10}$$

Un vector puede tener una longitud mayor debido a que tiene más valores distintos de 0, o porque cada dimensión tiene un valor más elevado. Cualquiera de estos casos puede incrementar el resultado del producto escalar. Por tanto, si consideramos el vector asociado a una palabra muy frecuente, tendrá un mayor número de valores distintos de 0 y tendrá valores mayores en cada dimensión (aunque se utilicen valores que controlen de alguna forma la frecuencia). Por tanto, el producto escalar favorece a las palabras más frecuentes.

Debido a este problema, es necesario modificar el producto escalar de forma que se normalicen los vectores y no se le dé más importancia a las palabras más frecuentes. Este producto escalar normalizado es el coseno del ángulo formado por los dos vectores. Esta medida ha sido utilizada frecuentemente en sistemas de Recuperación de Información (Frakes y Baeza-Yates (1992)) y ha

sido aplicada para calcular la similitud entre palabras a partir de sus relaciones gramaticales (Ruge (1992)).

La fórmula del coseno es la siguiente:

$$coseno(\overrightarrow{x}, \overrightarrow{y}) = \frac{\overrightarrow{x} \cdot \overrightarrow{y}}{|\overrightarrow{x}| \times |\overrightarrow{y}|} = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \times \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{5.11}$$

Gracias a la normalización de los vectores, la métrica del coseno no es sensible a valores extremos tal y como ocurría con la Manhattan y la Euclídea. El coseno abarca desde 1 para vectores muy próximos entre sí hasta 0 para vectores ortogonales que no tienen ninguna característica en común y -1 para vectores apuntando en direcciones opuestas (en la práctica los valores tienden a ser positivos).

Además de la medida del coseno, existen otras medidas de similitud muy utilizadas en el campo de PLN. Entre las medidas más representativas destacan: la medida de Kullback-Leibler (o entropía relativa) (Cover y Thomas (1991)), la medida de Jensen-Shannon (o radio de información) (Rao (1982), Lin (1991)) y el coeficiente de Jaccard (Jaccard (1901)). A continuación se definen cada una de ellas.

Universitat d'Alacant
Universidad de Alicante

$$KL(\overrightarrow{x}||\overrightarrow{y}) = \sum_{i=1}^{n} x_i \log \frac{x_i}{y_i} \tag{5.12}$$

La medida de Kullback-Leibler (KL) 5.12 determina la similitud entre dos distribuciones $x = \{x_i\}$ e $y = \{y_i\}$ y ha sido utilizada en (Dagan et al. (1994)) para medir la similitud entre palabras. Esta medida no es simétrica, de forma que, $KL(\overrightarrow{x}||\overrightarrow{y}) \neq KL(\overrightarrow{y}||\overrightarrow{x})$. Los valores que obtiene se encuentran entre 0 e infinito y únicamente adopta el valor de 0 cuando las dos distribuciones son exactamente iguales.

$$JS(\overrightarrow{x}, \overrightarrow{y}) = \frac{1}{2}[KL(\overrightarrow{x}||\frac{\overrightarrow{x} + \overrightarrow{y}}{2}) + KL(\overrightarrow{y}||\frac{\overrightarrow{x} + \overrightarrow{y}}{2})] \quad (5.13)$$

La medida de Jensen-Shannon (JS) 5.13, ha sido utilizada para medir la similitud entre palabras en (Dagan et al. (1997), Lee (1997)). Esta medida es la media de la suma de los dos valores de divergencia de KL entre las distribuciones x e y, tomando valores entre 0 y $2 \log 2$.

$$Jaccard(\overrightarrow{x}, \overrightarrow{y}) = \frac{|\{i : x_i > 0 \wedge y_i > 0\}|}{|\{i : x_i > 0 \vee y_i > 0\}|} \quad (5.14)$$

El coeficiente de Jaccard 5.14 mide la similitud entre dos distribuciones de datos. Dados dos conjuntos x e y el resultado de esta medida es la cardinalidad de su intersección dividida por la cardinalidad de su unión.

Además de las mencionadas anteriormente existen otras propuestas de métricas de similitud. Una comparación entre las diferentes métricas de similitud la encontramos en (Dagan et al. (1999)) y (Lee (1999)).

Por su simplicidad y adaptación a diferentes tamaños de los contextos, en este trabajo se ha optado por utilizar la métrica del coseno para evaluar la similitud entre vectores de características.

5.1.7 Determinación del sentido correcto

El objetivo del método DRelevant es determinar el sentido correcto de las palabras polisémicas que aparecen en un contexto determinado. Hasta ahora se han extraído los dominios relevantes a partir de WND y SUMO, se han definido los vectores de características que van a modelar el espacio semántico de palabras, frases, párrafos, etc y por último se ha establecido la métrica de similitud entre vectores. El último paso es la determinación del sentido correcto de cada palabra utilizando toda la información obtenida a partir del contexto y de WND y SUMO.

A continuación se va a explicar la mecánica del proceso mediante un ejemplo aplicado sobre WND.

Universitat d'Alacant
Universidad de Alicante

5.1.7.1 Ejemplo ilustrativo sobre WND.

Supongamos que queremos desambiguar la palabra *"image"* del siguiente texto extraído del British National Corpus (BNC):

*"A successful description of a self-portrait may not be difficult, but an illuminating interpretation may call on many references, especially other artists' pictures of themselves. What can be deduced from a self-portrait is often controversial; a critic is especially likely to read into a self-portrait some opinion held about the artist. When, as in the cases of Rembrandt and Van_Gogh, there is a whole series of pictures to choose from, books can be written on the self-**images** of one artist alone. This theme is a useful one for assessing the quality of a critic's writing, since it tempts the rash into speculation, while an impoverished eye will miss relevant and useful comparisons. A theme where personal psychology is necessarily absent is the Christian subject of the Madonna and Child."*

El primer paso, es construir el vector de contexto que modela semánticamente el contenido del párrafo donde aparece la palabra ambigua. Se utiliza un POS-tagger para obtener los lemas de todas las palabras y extraer sus correspondientes dominios relevantes. El resultado es un vector de características obtenido a partir de los dominios relevantes de todas las palabras del contexto.

En la Figura 5.8 se muestran los lemas obtenidos a partir del contexto. Y en la Figura 5.9 se muestra el vector de contexto obtenido a partir de los Dominios Relevantes.

El segundo paso, es obtener un segundo vector con el que establecer el grado de similitud con el vector de contexto extraído anteriormente. Dado que se desea obtener el sentido correcto de la palabra *"image"*, es necesario disponer de vectores de características que modelen cada uno de los posibles sentidos de esta palabra. Para ello, se utilizan las glosas de WordNet. Es decir, para cada uno de los sentidos de *"image"* se extrae su glosa y se construye un vector de características, denominado en este caso vector de sentido.

NOMBRES
description self portrait interpretation reference artist picture opinion case Van Gogh series book ***image*** *theme one quality critic writing rash speculation eye comparison writing rash psychology subject Madonna Child*
VERBOS
be call deduce read hold choose write assess tempt miss
ADJETIVOS
successful difficult illuminating many other controversial likely whole useful impoverished relevant personal absent Christian

Figura 5.8. Lemas del contexto

{ doctrines 0.000264501
art 0.000252649
history 0.00016812
heraldry 0.000243991
linguistics 0.000148377
grammar 0.000276004
literature 0.000195926
philology 0.00118873
philosophy 0.000871316
... }

Figura 5.9. Vector de contexto

En la Tabla 5.7 se muestran los 7 sentidos para *"image"* con sus respectivas glosas.

En la Figura 5.10 se muestran los vectores de sentido para los 3 primeros sentidos de *"image"*.

El tercer paso, una vez obtenidos los distintos vectores de sentido, es medir el grado de similitud entre cada uno de estos vectores y el vector de contexto. Aquel vector de sentido cuyo coseno con el vector de contexto obtenga el mayor valor, será el elegido como el sentido correcto de la palabra *"image"*.

En este ejemplo, el sentido correcto para *"image"* es el 2. Tal y como muestra la Figura 5.11, el resultado del coseno entre el vector de contexto y cada uno de los vectores de sentido demuestra que el sentido más adecuado en este caso sería el 2.

Synset	Dominio	Sentido	Glosa
04551473	factotum	image#1	an iconic mental representation; "her imagination forced images upon her too awful to contemplate"
03685960	psychology	image#2	(Jungian psychology) a personal facade one presents to the world; "a public image is as fragile...
03118233	factotum	image#3	a visual representation of an object or scene or person produced on a surface; "they showed us the...
04559702	factotum	image#4	a standard or typical example; "he is the prototype of good breeding"; "he provided America with a...
05317127	literature	image#5	language used in a figurative or... nonliteral sense
07223613	person	image#6	someone who closely resembles a famous person (especially an actor); "he could be Gingrich's double"...
02622723	factotum	image#7	a likeness of a person (especially in the form of sculpture); "the coin bears an effigy of Lincoln...

Tabla 5.7. Glosas para *"image"*

image#1	**image#2**	**image#3**
rowing 0.006974	plastic_arts 0.003155	plastic_arts 0.002746
table_tennis 0.003351	wrestling 0.003151	applied_science 0.001671
photography 0.003125	cycling 0.003146	astrology 0.001606
psychiatry 0.002593	astrology 0.002747	electrotechnics 0.001547
radiology 0.001896	surgery 0.002375	radiology 0.001544
tv 0.001804	post 0.002209	cinema 0.001338
statistics 0.001599	applied_science 0.001851	textiles 0.001331
auto 0.001161	electrotechnics 0.001670	tv 0.001205
tennis 0.001119	number 0.001574	sculpture 0.001091
...	...	...

Figura 5.10. Vectores de sentido para *"image"*

En la Figura 5.12 se muestra gráficamente todo el proceso necesario para obtener el sentido de una palabra a partir del contexto que la rodea. Este proceso requiere el establecimiento en primer lugar de la cantidad de palabras que van a formar parte del contexto (ventana de N palabras, párrafo, oración...) y a partir de ahí comienza la tarea de desambiguación.

04551473—factotum —image#1— 0.161629
03685960—psychology —image#2— 0.79989
03118233—factotum —image#3— 0.485136
04559702—factotum —image#4— 0.54842
05317127—literature —image#5—0.111334
07223613—person —image#6— 0.690172
02622723—factotum —image#7— 0.149694

Figura 5.11. Resultado del coseno entre VC y VS's

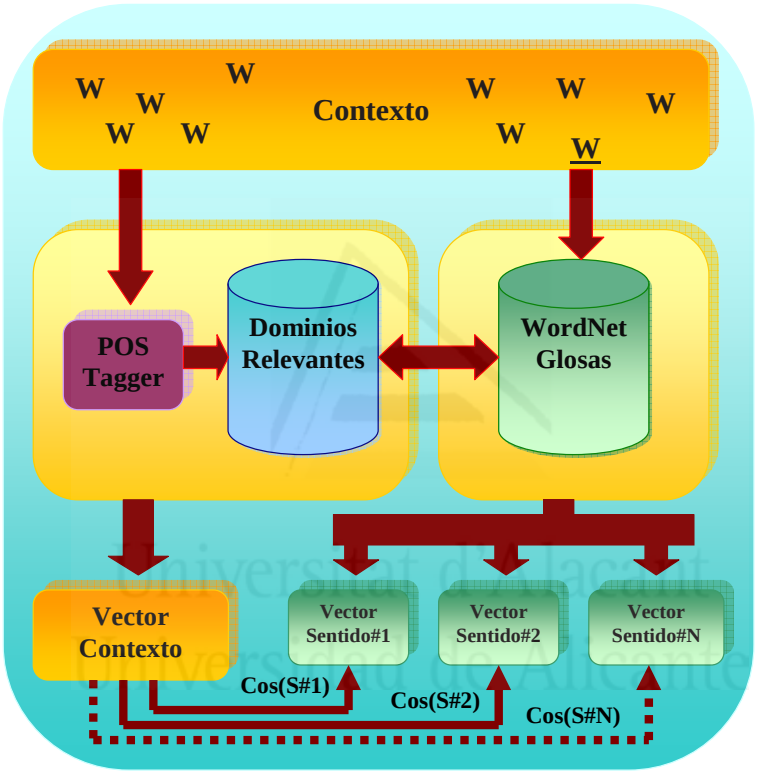


Figura 5.12. Sistema DRelevant

5.1.8 Extended WordNet y Dominios Relevantes

En el proceso descrito anteriormente para la obtención de los Dominios Relevantes a partir de las glosas de WordNet Domains, la hipótesis fundamental era que las palabras de la glosa que definen un sentido de una palabra están relacionadas semánticamente, y por tanto, es muy probable que compartan el mismo dominio. Como resultado, usando la información proporcionada por el dominio asociado al sentido de la palabra, se etiquetaron las palabras de la glosa y se extrajeron los Dominios Relevantes para cada palabra de WordNet. Sin embargo, este proceso podría ser mejorado si se tuvieran en cuenta los sentidos correctos de las palabras de las glosas, ya que, de esta forma la anotación sería mucho más precisa.

Dada la gran cantidad de aplicaciones actuales que utilizan WordNet como fuente de información: sistemas de WSD, RI, QA, etc, que requieren de información lógica y semántica adicional (de la que WordNet carece), se ha planteado la mejora de este recurso mediante la incorporación de información sintáctica, lógica y semántica, en el denominado Extended WordNet (Harabagiu et al. (1999), Mihalcea y Moldovan (2001)). Este nuevo recurso ha incorporado información adicional a las glosas de WordNet, de forma que cada palabra de las glosas viene acompañada de información sintáctica, lema, sentido, etc. Un ejemplo de esta nueva anotación se muestra en la Figura 5.13.

Según muestra la Figura 5.13, hay tres fuentes de información añadidas a la glosa inicial:

- **Información semántica.** Comprendida entre las etiquetas <wsd> </wsd>. A cada palabra de la glosa se le asigna su correspondiente sentido, utilizando para ello un proceso semiautomático de desambiguación.
- **Información sintáctica.** Comprendida entre las etiquetas <parse> </parse>. Se extraen las categorías morfosintácticas de cada elemento que conforma la glosa.
- **Información lógica.** Comprendida entre las etiquetas <lft> </lft>. Se transforma cada glosa en su forma lógica correspondiente.

```
<gloss pos="NOUN" synsetID="09786238">
 <synonymSet>president</synonymSet>
<text>
 the chief executive of a republic
</text>
 <wsd>
    <wf pos="DT" >the</wf>
    <wf pos="NN" lemma="chief_executive" quality="normal" wnsn="2" >
                                                        chief_executive</wf>
    <wf pos="IN" >of</wf>
    <wf pos="DT" >a</wf>
    <wf pos="NN" lemma="republic" quality="normal" wnsn="1" >republic</wf>
 </wsd>
<parse quality="SILVER">
(TOP (S (NP (NN president) )
      (VP (VBZ is)
         (NP (NP (DT the) (JJ chief) (NN executive) )
            (PP (IN of)
               (NP (DT a) (NN republic) ) ) ) )
      (. .) ) )
</parse>
 <lft quality="GOLD">
  president:NN(x1) -> chief_executive:NN(x1) of:IN(x1, x2) republic:NN(x2)
 </lft>
</gloss>
```

Figura 5.13. Extended WordNet para president#3

Universitat d'Alacant

En nuestro caso, sólo vamos a utilizar la información semántica comprendida entre las etiquetas <wsd> </wsd>, para poder realizar una anotación más precisa de los dominios de las glosas de WordNet. Así pues, utilizando los sentidos de las palabras de las glosas, se extraen sus dominios asociados. De este modo, el recurso léxico Dominios Relevantes mejorará la calidad de anotación de las palabras de WordNet.

Un ejemplo práctico que demuestra la diferencia de anotación de WordNet Domains frente a Extended WordNet, es el mostrado en la Figura 5.14.

En este caso, se tiene la glosa asociada al sentido 3 de la palabra *"president"*, cuyos dominios asociados son PERSON y POLITICS. En la parte de la izquierda se muestra la anotación de dominios

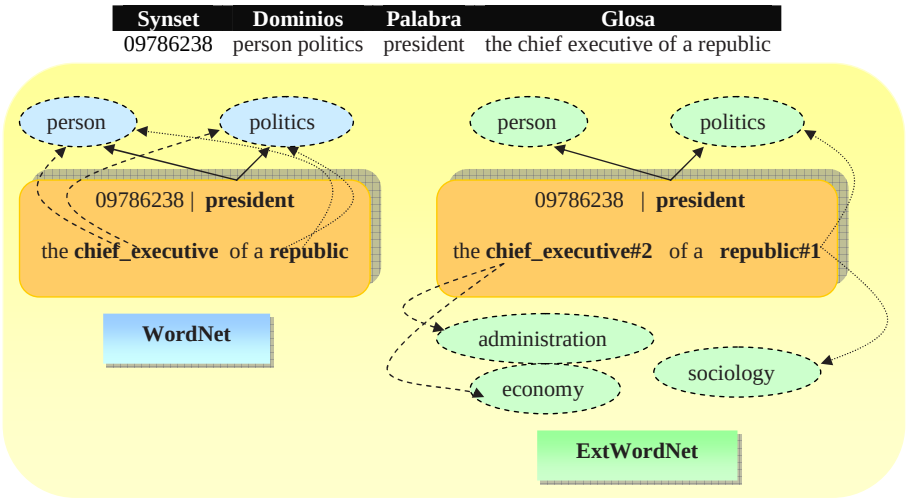


Figura 5.14. Extracción de dominios con Extended WordNet

mediante WordNet Domains, cuyo resultado es la asignación de los dominios PERSON y POLITICS a todas las palabras con contenido semántico de la glosa. En la parte derecha se muestra el resultado de anotación tras utilizar Extended WordNet. En este caso, al proporcionar el sentido correcto de las palabras de la glosa, se pueden extraer los dominos asociados a ellas, mejorando de esta forma la anotación.

En la Figura 5.15 se muestra un ejemplo comparativo de anotación con respecto a la palabra *"president"* usando WordNet Domains y Extended WordNet.

Tal y como muestra la Figura 5.15, los Dominios Relevantes para *"president"* difieren dependiendo de la fuente de información utilizada. En (Vázquez et al. (2007))se compararon los resultados obtenidos por el método de desambiguación DRelevant usando las dos alternativas. El análisis posterior determinó que se obtenían mejores resultados al utilizar los Dominios Relevantes enriquecidos con la información de las glosas de Extended WordNet.

La evaluación del método DRelevant se ha realizado sobre las tareas *"All Words"* y *"Lexical Sample"* de SENSEVAL. Además, el recurso léxico Dominios Relevantes ha sido empleado sobre otras

DR WND	DR ExtWN
administration 0.001588	politics 0.070828
building_industry 0.000389	history 0.067804
art 0.000366	enterprise 0.002493
commerce 0.000420	military 0.001694
economy 0.001077	economy 0.001420
enterprise 0.000628	university 0.001280
factotum 0.000385	telecommunication 0.001030
geography 0.000321	administration 0.000949
history 0.080090	art 0.000795
...	...

Figura 5.15. Dominios Relevantes (DR) para *"president"*

tareas de PLN como la resolución de la implicación textual. En el Capítulo de Evaluación se muestran los resultados obtenidos.

5.2 WSD basado en conocimiento: DLSA

En el capítulo anterior se hacía referencia a un modelo computacional denominado LSA (Análisis de la Semántica Latente). La idea de utilizar este modelo computacional para tratar el problema de la ambigüedad semántica se debe a la capacidad que tiene LSA de adaptarse a cualquier idioma, contexto o circunstancia. Es decir, LSA es una técnica basada en modelos estocásticos del significado o modelos semánticos del lenguaje, la cual, no da importancia a las categorías léxicas, orden de las palabras, conjugaciones verbales, etc.

Aunque inicialmente LSA fue concebida como una técnica para Recuperación de Información (Deerwester et al. (1990)), fue más tarde en (Landauer y Dumais (1997)) cuando se propuso la utilización de esta técnica para la adquisición y representación del conocimiento. A partir de entonces LSA ha sido aplicada a diferentes áreas: corrección de textos en el ámbito académico (Haley et al. (2005)), cohesión y coherencia de textos (Graesser et al. (2004)), complemento de ontologías (Cederberg y Widdows (2003)), categorización de documentos (Rosso et al. (2004)), etc.

En cuanto a la tarea de WSD, LSA ya ha sido utilizada previamente, de forma que a partir de un algoritmo no supervisado se inducen las similitudes entre palabras, basándose en co-ocurrencias entre términos y contextos. Tal y como se describió en el capítulo anterior, la dimensión del espacio vectorial inicial de la matriz se reduce utilizando la descomposición en valores singulares. Finalmente, mediante la utilización de técnicas de clustering, se identifican los sentidos de las palabras (Schütze (1998), Widdows y Peters (2003)).

En nuestro caso de estudio, LSA se utiliza para extraer información semántica a partir de la información contextual y la co-ocurrencia de palabras. Posteriormente, la información proporcionada por LSA es utilizada como fuente de conocimiento para el nuevo método de desambiguación automática DLSA.

5.2.1 Base de datos léxica como fuente de conocimiento

A medida que aumenta la cantidad de documentos que intervienen en la matriz original de LSA, también aumenta la posibilidad de que un mismo término aparezca en documentos distintos con el mismo significado (palabras polisémicas). Esta circunstancia produce como consecuencia la introducción de ruido y confusión en el espacio conceptual obtenido, ya que, cada documento se considera independiente del resto. Una solución a este problema sería la clasificación inicial de los documentos a partir de diferentes tópicos, para así mitigar los efectos de la distribución de palabras bajo el mismo sentido en diferentes documentos. Sin embargo, no existen actualmente sistemas lo suficientemente precisos como para realizar esta clasificación de forma automática, sin un costo computacional excesivo.

Además del problema de la distribución de palabras polisémicas existe otro problema asociado a la utilización de documentos como fuente de información: el tópico tratado. Es decir, si para la obtención de datos se utilizan documentos relacionados, por ejemplo, con medicina o biología, probablemente la palabra “planta” estará relacionada con palabras como: “perenne”, “fotosíntesis”, “polen”, etc. Quedando relegados el resto de sentidos de la pala-

bra "planta" en favor del sentido asociado al tópico o registro de los documentos. Por ello, para paliar los efectos de la utilización de documentos como fuente de información, es necesario utilizar otro tipo de recurso que de alguna forma distribuya según criterios semánticos las palabras polisémicas y sea independiente del dominio.

Para solucionar los problemas derivados de la utilización de documentos como fuente de información, es necesario utilizar un recurso que agrupe conceptos relacionados semánticamente bajo una serie de categorías genéricas. Nuestra propuesta es la utilización de las categorías semánticas de WND o SUMO, como fuente de conocimiento que agrupa palabras relacionadas semánticamente y a partir de las cuales se puede construir una matriz conceptual. De esta forma, se logra una matriz conceptual independiente del dominio, ya que, se construye sobre una base de datos léxica que utiliza los conceptos de forma genérica, donde cada aparición de una palabra bajo una categoría semántica (dominio) implica la asociación de un determinado sentido. Por tanto, se evita por una parte la dependencia del dominio que se daba en el caso de los documentos y además se soluciona el problema de la distribución de los mismos sentidos en diferentes documentos.

5.2.2 LSA aplicado a WSD

La utilización de categorías semánticas como base para modelar los diferentes espacios contextuales de la matriz, proporciona información muy útil acerca de los sentidos de las palabras. Así pues, a partir de los valores de similitud obtenidos tras aplicar LSA sobre palabras, oraciones, o párrafos, se puede establecer a qué dominio pertenece un determinado contexto. De esta forma, si se intenta establecer el significado de una palabra dentro de una frase f_1 se realiza el cálculo sobre la matriz conceptual, obteniendo un listado con los dominios más significativos en relación a esa frase (en lugar de los documentos más similares). De este modo, es posible establecer el concepto semántico sobre el que subyace un determinado contexto, y a partir de ahí se pueden determinar los sentidos de las palabras que lo componen.

El método que se propone en esta sección utiliza como fuente de conocimiento la base de datos léxica WND y el recurso léxico Dominios Relevantes. El proceso es el siguiente: a partir de la información proporcionada por las glosas de WordNet Domains se construye la matriz conceptual, donde las columnas se corresponden con los dominios de la jerarquía de WND y las filas son las palabras de las glosas. Previamente a la obtención de la matriz se realiza la lematización de todas las palabras, ya que, LSA considera conceptos distintos, palabras en plural o en singular, o las diferentes conjugaciones verbales. Los experimentos realizados en (Vázquez et al. (2006)), donde se utiliza la técnica de LSA para determinar la implicación textual, demuestran que tras la lematización de las palabras que conforman la matriz conceptual los resultados mejoran.

Una vez obtenida la matriz conceptual se realiza su descomposición en valores singulares reduciendo las dimensiones de la matriz de 162 a 100. De esta forma, los dominios que están dentro de una subjerarquía se agrupan y así se consigue mantener una mejor cohesión semántica en la matriz conceptual.

Tras la reducción de dimensiones ya es posible determinar a qué categorías semánticas está asociada una oración, un párrafo, etc. Así pues, se realiza la comparación entre el vector de palabras previamente lematizadas del contexto de entrada con la matriz conceptual. El resultado es un listado con las categorías a las que pertenece el contexto de entrada junto con su grado de similitud.

A partir de las categorías (dominios) obtenidas como resultado de la aplicación de LSA, se han desarrollado una serie de heurísticas para determinar el sentido más apropiado de las palabras del contexto. Todas las heurísticas determinan el sentido de las palabras del contexto inicial utilizando como fuente de información los dominios más significativos proporcionados por LSA.

5.2.2.1 Heurística 1: Ratio de Asociación.

Esta heurística utiliza los valores del Ratio de Asociación de los Dominios Relevantes como base para determinar el sentido de las palabras. En este caso, se extraen los 10 o 20 primeros domi-

nios a partir de LSA. Estos dominios conforman un conjunto que se intersecciona con el conjunto de Dominios Relevantes de cada uno de los posibles sentidos de la palabra ambigua. De esta forma, para cada posible sentido, se seleccionan únicamente los dominios compartidos con el resultado de LSA. El sentido seleccionado es aquel con el valor más elevado para los diferentes valores de RA obtenidos.

5.2.2.2 Heurística 2: Similitud LSA.

En esta heurística únicamente se computan los valores de similitud obtenidos por LSA. Cada dominio devuelto por LSA tiene asociado un valor de similitud. En este caso, si el dominio devuelto por LSA se encuentra entre los dominios relevantes de un determinado sentido de la palabra, se almacena su valor de similitud. El resultado es el sumatorio de todos aquellos valores de similitud cuyos dominios se encuentran entre el listado de los dominios relevantes de cada uno de los sentidos. El sentido elegido es aquel cuyo sumatorio es mayor.

5.2.2.3 Heurística 3: Similitud LSA × Ratio de Asociación.

Esta última heurística determina el sentido de las palabras a partir del valor obtenido tras multiplicar el valor de similitud obtenido por LSA por el RA de los dominios de cada uno de los sentidos. En este caso, se le da preferencia a aquellos dominios cuyo valor de similitud obtenido por LSA es mayor. De esta forma, se favorecen los dominios con mayor peso semántico sobre la matriz conceptual. Como en los casos anteriores el sentido seleccionado es aquel con mayor valor.

5.2.2.4 Ejemplo ilustrativo.

Supongamos que tenemos el siguiente texto del que queremos desambiguar el verbo “add”:

Universitat d'Alacant

"The two Roman catholic priests, who were in all respects dedicated pastors and much liked by many in the local community, immediately opposed the idea, preaching against it at Sunday masses in the local convent and the school hall. The burden of the message was that good catholic parents sent their children to catholic schools. The curate ***added*** *to this that those promoting the integrated project were in fact promoting secularism. The fact that a prominent member of the current community council and an integrated education supporter was a member of official Sinn Fein, the Workers' party, appeared to figure in the reasoning, as this party has always been suspected to be an anti - clerical and secularist force. In residents' association meetings, the clergy's point of view received vocal support from one or two members of the older village community which preceded the housing estate."*

El primer paso es obtener los lemas de todas las palabras con contenido semántico (nombres, verbos adjetivos y adverbios). Una vez obtenidos los lemas se establece el grado de similitud entre el vector del contexto donde está la palabra ambigua con respecto a la matriz conceptual de LSA (ver Figura 5.16).

En este caso, el resultado obtenido es:

{
theatre 0.800681
music 0.796623
play 0.795698
skiing 0.786670
optics 0.776752
badminton 0.775749
basketball 0.767051
hockey 0.763763
card 0.762485
sport 0.733817
...

Figura 5.16. Dominios Relevantes según LSA

Con esta información se procede a realizar la intersección con cada uno de los sentidos del verbo *"add"*. Los tres primeros vectores de sentidos son los mostrados en la Figura 5.17.

Sentido 1	Sentido 2	Sentido 3
post 0.003351	oceanography 0.004109	badminton 0.003255
table_tennis 0.001158	numismatics 0.003551	wrestling 0.002423
plastic_arts 0.000811	basketball 0.003037	plastic_arts 0.002387
tennis 0.000750	tax 0.002671	cycling 0.002387
banking 0.000695	occultism 0.001521	surgery 0.001844
artisanship 0.000641	exchange 0.001425	football 0.001549
wrestling 0.000593	archery 0.000855	applied_science 0.001404
philately 0.000590	diving 0.000784	cinema 0.001309
school 0.000504	gas 0.000708	university 0.001072
ethnology 0.000483	psychoanalysis 0.000699	electrotechnics 0.001055
...	...	...

Figura 5.17. Vectores de sentidos para *"add"*

Tras la comparativa y extracción de los dominios presentes en la intersección de cada vector de sentido con el vector de contexto obtenido a partir de LSA, el resultado utilizando la heurística 2, es el mostrado en la Figura 5.18.

add#1 - 9.644582
add#2 - 7.703261
add#3 - 6.741430
add#4 - 6.646582
add#5 - 6.737763
add#6 - 5.798008

Figura 5.18. Selección del sentido correcto

Universitat d'Alacant
Universidad de Alicante

De tal forma, que el sentido seleccionado finalmente como el apropiado para *"add"* es el sentido 1.

5.2.3 LSA aplicado a NED

Un campo que actualmente está teniendo una gran repercusión es el relacionado con la detección y discriminación de entidades: nombres propios, lugares, organizaciones, etc.

La demanda de sistemas que por ejemplo, clasifiquen páginas web referentes a una determinada persona o una organización,

es cada vez más elevada. Esta necesidad surge debido a la gran proliferación de Internet en los últimos años, la aparición de blogs, páginas personales, etc.

Al igual que existen palabras polisémicas encontramos entidades nombradas que comparten la misma forma nominal pero difieren en significado. Por ejemplo, las siglas ACM pertenecen a: Association for Computing Machinery, Associació Catalana de Municipis i Comarques, Actividades de Carpintería de Madera, Asociación de Celíacos de Madrid, etc. En este caso, un sistema de clasificación y discriminación debería distinguir entre las distintas acepciones de ACM, basándose primordialmente en los contextos de cada página. Es por ello, que hemos considerado la idea de aprovechar las ventajas que proporciona LSA para determinar la similitud entre contextos y utilizarla para extraer aquellas páginas web o documentos que hagan referencia a una misma entidad.

En el apartado de Evaluación se muestran los resultados tras utilizar la técnica de LSA para desambiguación y discriminación de nombres propios.

5.3 WSD basado en reglas lingüísticas sobre corpus

Uno de los principales problemas en la resolución automática de la ambigüedad semántica, radica en la falta de utilización de conocimiento lingüístico (Manning y Schütze (1999), Calzolari et al. (2001), etc). En la actualidad, la gran mayoría de los sistemas se centran en el desarrollo de complejos algoritmos estadísticos que manejan muy poca cantidad de información lingüística. Considerando esta deficiencia, se hace por tanto necesaria la incorporación de más información lingüística que revele ciertas propiedades y relaciones en el lenguaje, no evidenciadas a través de estimaciones estadísticas.

El principal objetivo del método desarrollado en esta sección es demostrar que mediante la utilización de información lingüística se obtiene un elevado porcentaje de aciertos en términos de precisión. Este incremento de la precisión obtenida viene acom-

pañado de un decremento de la cobertura, debido en gran parte a la escasez de recursos lingüísticos (en español, en nuestro caso), que impiden la obtención de información lingüística a gran escala.

A continuación se exponen las características de este método junto con algunos ejemplos aplicados a WSD.

5.3.1 Obtención de información lingüística

La decisión de utilizar información lingüística para el desarrollo de un método de desambiguación automática, es debida a la necesidad de evitar, en la medida de lo posible, el uso de parámetros estadísticos, frecuencias, medidas de similitud, etc, que llevan asociados un desvío y un margen de error, resultado de tratar los textos desde el punto de vista matemático.

El método presentado en esta sección utiliza información implícita presente en textos no anotados semánticamente como base para la adquisición de conocimiento. De esta forma, no es necesaria una anotación manual previa del corpus.

5.3.1.1 Adquisición de información paradigmática.

Actualmente, la obtención de información sintagmática relacionada con los sentidos de las palabras es difícil de obtener y muy costosa. El término información sintagmática hace referencia a palabras que co-ocurren frecuentemente, incluyendo colocaciones y restricciones de selección. Sin embargo, la información paradigmática, que hace referencia a palabras que aparecen en contextos similares (hipónimos/hiperónimos, merónimos/holónimos, etc), es más sencilla de obtener.

En esta sección se va a describir cómo la obtención de información paradigmática ayuda a la extracción de información sintagmática. Esta idea está basada en la hipótesis de que palabras similares semánticamente (eje paradigmático) pueden ser sustituidas en el mismo contexto (eje sintagmático). Tal y como muestra la Figura 5.19:

En este caso, si se fija la secuencia *"obra para órgano"*, y se deja libre la posición ocupada por la palabra *"obra"*, se pueden

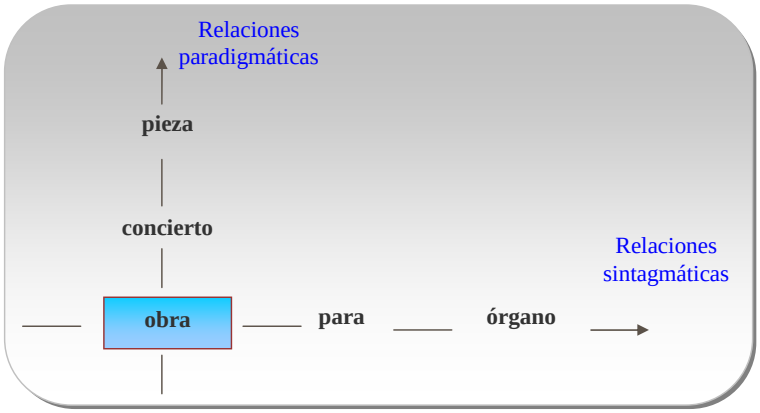


Figura 5.19. Relaciones sintagmáticas y paradigmáticas

extraer otras palabras similares semánticamente que pueden sustituir a *"obra"*, como por ejemplo: *"concierto"*, *"pieza"*, *"composición"*, etc. Estas nuevas palabras pueden intercambiarse entre sí, de forma que el contexto mantiene su significado original: *"composición musical para un instrumento"*. Según el estudio realizado en (Miller y Charles (1991)), se demuestra que un individuo determina la similitud semántica entre palabras, tomando como base los contextos en los que se utilizan.

Universitat d'Alacant
Universidad de Alicante

5.3.1.2 Discriminadores de sentidos.

A partir de la información paradigmática proporcionada, se puede establecer el sentido de la palabra ambigua. Es decir, a partir de todas las palabras que conforman el eje paradigmático, que pueden ponerse en lugar de la palabra ambigua, y usando una fuente de datos léxica que proporcione información para cada uno de los posibles sentidos de la palabra, se puede establecer el sentido más adecuado de la palabra ambigua.

En nuestro caso, la fuente de datos léxica elegida para la obtención de los diferentes sentidos de las palabras, ha sido EuroWordNet[1] (Vossen (1998)). EuroWordNet es una extensión multilingüe

[1] `http://www.globalwordnet.org/`

de WordNet para 8 idiomas europeos (inglés, español, alemán, holandés, italiano, francés, estonio y checo). Cada synset de los diferentes idiomas se encuentra conectado con el resto, a través del llamado *"Inter Lingual Index"* (ILI). En la Figura 5.20 se muestra la interconexión entre los diferentes idiomas.

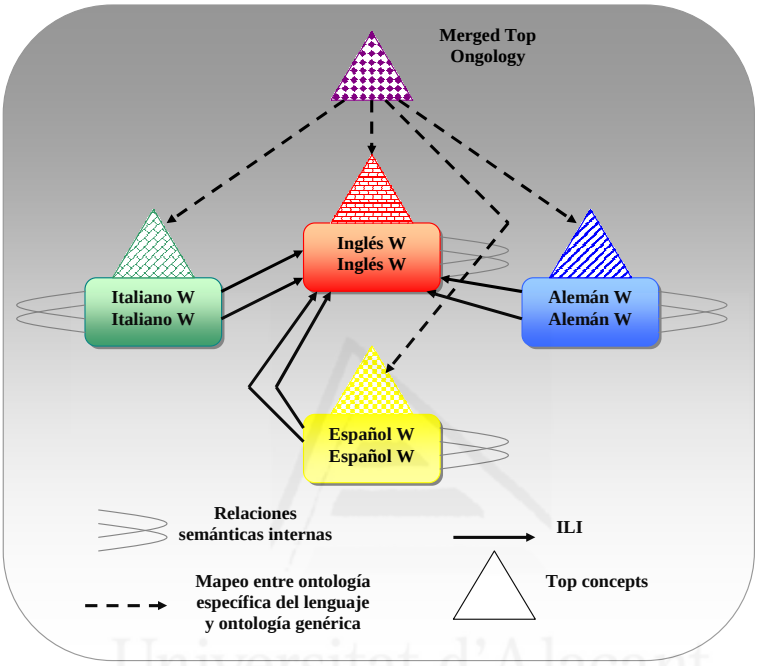


Figura 5.20. EuroWordNet

Universitat d'Alacant
Universidad de Alicante

De esta forma, utilizando EuroWordNet y sus relaciones semánticas, el método desarrollado en esta sección se va a aplicar sobre el idioma español. Para ello, se va a emplear el conjunto de *"variants"*[2] proporcionados para cada uno de los sentidos de una palabra. Así pues, para cada sentido W_i de una palabra W, se extraen de EWN el conjunto de synsets con los que se relaciona (hipónimos, hiperónimos, merónimos, etc), junto con los *"va-*

[2] Los *"variants"* son las palabras que forman un synset. Por ejemplo, los *"variants"* del synset número 00589353 de EWN son: tranquilidad, solaz, reposo, relax, relajo, huelga...

riants" que los identifican. El resultado final es la asignación al sentido W_i todos los *"variants"* V_i extraídos de las relaciones léxico semánticas de EWN. Una vez obtenidos los conjuntos V_i para cada sentido i, el siguiente paso es eliminar aquellos *"variants"* que se encuentren en más de un sentido, obteniendo de esta forma una serie de conjuntos disjuntos D_i con palabras relacionadas exclusivamente con el sentido W_i de la palabra. El resultado final es una serie de conjuntos D_i denominados de aquí en adelante como Discriminadores de Sentidos.

A continuación se muestra el proceso de extracción de los Discriminadores de Sentidos para la palabra *"órgano"* que según EuroWordNet tiene cinco sentidos diferentes:

órgano#1: "parte de una planta"
órgano#2: "agencia gubernamental, instrumento"
órgano#3: "parte funcional de un animal"
órgano#4: "instrumento musical"
órgano#5: "periódico"

Partiendo de órgano#1 extraemos sus *"variants"* a través de las relaciones léxico semánticas:

Sinonimia: órgano vegetal.
Hiperonimia: objeto inanimado, objeto físico, objeto, cosa, entidad.
Hiponimia: estructura reproductiva, lámina, raíz, tronco, tallo, rabillo, pedúnculo, cálamo, caña, cabillo, hoja, follaje, ...
Holonimia: $\oslash$.
Meronimia: $\oslash$.
Coordinados: talo, sombrero, sombrerito, sombrerillo, sombrerete, píleo, carpóforo, volva, chalaza,...

Una vez obtenidos todos los *"variants"* se construye el conjunto V_1={órgano vegetal, objeto inanimado, objeto físico, objeto, cosa, entidad, estructura reproductiva, lámina, raíz, tronco, tallo, rabillo,...}. Y así para cada uno de los sentidos de "órgano". Finalmente, se extraen los Discriminadores de Sentidos, quedando el conjunto D_1={órgano vegetal, lámina, raíz, tronco, tallo,...}.

Para abarcar más información derivada de las relaciones de EuroWordNet, el conjunto de Discriminadores de Sentidos se ha

ampliado utilizando las relaciones heredadas de hiponimia, manteniendo la característica de conjuntos disjuntos. De esta forma, a estos nuevos conjuntos ampliados se les denomina DE_i (Discriminadores de Sentidos Extendidos).

La obtención del sentido de una palabra viene determinado por el número de palabras del eje paradigmático que tiene en común con cada uno de los conjuntos disjuntos de Discriminadores de Sentidos. A mayor número de palabras en común con el conjunto DE_i, mayor probabilidad de que la palabra W tenga asociado el sentido i.

5.3.1.3 Identificación de patrones sintagmáticos.

Para la determinación de la información paradigmática es necesario el establecimiento, en primer lugar, de una serie de patrones sintagmáticos que identifiquen entidades con contenido semántico susceptibles de ser intercambiadas con otras entidades en el mismo contexto (eje sintagmático). De esta forma, cobra importancia la determinación del contexto local para cada ocurrencia y categoría sintáctica.

La aparición de una palabra en un contexto determinado proporciona información muy útil para su desambiguación (Ravin y Leacock (2001)). Asimismo, existe una interdependencia entre el significado de una palabra y el significado de las estructuras sintácticas superiores que la contienen: sintagma y oración. Un estudio realizado en (Miller y Charles (1991)) demuestra que una ventana de pocas palabras alrededor de la ocurrencia ambigua, es suficiente para obtener su significado. De esta forma, nuestra aproximación parte de un contexto mínimo para identificar el sentido de la ocurrencia ambigua, para posteriormente ampliarlo hasta obtener un único sentido posible.

Se parte por tanto, de una secuencia sintagmática reducida que contiene la ocurrencia ambigua, manteniendo fijos los demás elementos y se deja libre la posición que ocupa la palabra ambigua. De esta forma, se buscan en el corpus palabras que ocupen su lugar manteniendo la misma información sintagmática original (Ver Figura 5.19). A partir de aquí, se introduce un nuevo concepto, el de

patrones sintagmáticos, que son una tripleta formada por dos unidades lingüísticas $L1$ y $L2$ de contenido léxico (nombres, adjetivos, verbos o adverbios) y un patrón léxico-sintáctico R que expresa la relación (de dependencia, de coordinación, léxico-semántica o de adyacencia) que comparten las dos unidades léxicas. La relación R puede contener valores nulos, por ejemplo, en el caso de relaciones de adyacencia entre nombres y adjetivos:

pasaje**-L1** $\oslash$**-R** subterráneo**-L2**

Definimos dos tipos de patrones sintagmáticos:

1. Patrones sintagmáticos que corresponden a relaciones sintácticas (patrones sintácticos). Por ejemplo, corona de santo.
2. Patrones sintagmáticos que corresponden a relaciones léxico-semánticas (patrones léxico-semánticos). Por ejemplo, los miembros del comité (relaciones de meronimia).

Ambos tipos de patrones son relevantes para la identificación del sentido de la palabra ambigua. En el caso de los patrones sintácticos, el elemento relacional **R** se expresa mediante palabras funcionales, mientras que en los patrones léxico-semánticos **R** suele tener una forma más compleja y puede contener tanto palabras funcionales como de contenido léxico. En este trabajo, nos vamos a centrar en los patrones sintácticos, dejando como trabajo futuro la incorporación de patrones léxico-semánticos. Además, vamos a centrar el proceso de desambiguación sobre los nombres, por ser esta categoría la más rica en cuanto a relaciones sintácticas en sus proximidades.

Universidad de Alicante

Las hipótesis para determinar el sentido de una palabra ambigua a partir de patrones sintagmáticos son las siguientes:

1. Dos palabras que comparten un mismo patrón sintagmático, tienen una alta probabilidad de estar relacionadas semánticamente.
2. Dos ocurrencias de una palabra ambigua tienen una alta probabilidad de pertenecer al mismo sentido si aparecen en un mismo patrón sintagmático.

Para la identificación de patrones sintácticos se tienen en cuenta dos criterios:

Estructural (sintáctico). Se tienen en cuenta determinadas combinaciones de categorías morfosintácticas mediante las que se establecen relaciones sintácticas. En este caso, los patrones considerados son los siguientes:

- N1 (((ADV) ADV) ADJ/VPART), (PREP) (DET) (((ADV) ADV) ADJ/VPART) N2
- N1 (((ADV) ADV) ADJ/VPART) CONJ* (DET) (((ADV) ADV) ADJ/VPART) N2
- N ((ADV) ADV) ADJ/VPART (CONJ* ((ADV) ADV) ADJ/VPART)
- N1 (((ADV) ADV) ADJ/VPART) PREP (DET) (((ADV) ADV) ADJ/VPART) N2

Las categorías entre paréntesis hacen referencia a elementos que pueden estar o no presentes en el patrón. Y las categorías separadas por una barra son alternativas para una misma posición. Los patrones compuestos pueden descomponerse en patrones simples mediante unas reglas de descomposición predefinidas. Por ejemplo, el patrón [**N ADJ1 CONJ ADJ2**] se descompone en [**N ADJ1**] y [**N ADJ2**].

Frecuencia. Aquellos patrones que cumplen el criterio estructural se filtran de acuerdo a su frecuencia de aparición en el corpus.

Mediante estos criterios se eliminan aquellas combinaciones inaceptables de categorías sintácticas y las combinaciones poco frecuentes.

Una vez detectado el patrón sintáctico en el que aparece la palabra ambigua y las palabras relacionadas en el eje paradigmático, se puede establecer el sentido de la palabra ambigua utilizando los Discriminadores de Sentidos a partir de EWN.

5.3.2 Prueba de conmutabilidad

El algoritmo utilizado para obtener el sentido correcto de una ocurrencia ambigua es el denominado Prueba de Conmutabilidad. Este algoritmo utiliza la adaptación de EWN mediante Discriminadores de Sentidos para determinar el sentido correcto de una palabra.

Este algoritmo está basado en la hipótesis de que dos palabras que pueden conmutar en un contexto determinado están relacionadas semánticamente. En términos de los Discriminadores de Sentidos definidos anteriormente, si una palabra ambigua puede sustituirse en sus patrones sintácticos por un Discriminador de Sentido, entonces se le puede asignar el sentido correspondiente a ese Discriminador.

Para el ejemplo de la Figura 5.19, *"obra para órgano"*, si se quiere desambiguar la palabra "órgano", se deben buscan ocurrencias en el corpus del patrón [obra - para - X], donde X puede sustituirse por: violín, guitarra, piano, etc. Los nombres que pueden sustituir a *"órgano"* pertenecen al conjunto de Discriminadores de Sentido de órgano#4, y por tanto, *"órgano"* podría tener el sentido 4. La Figura 5.21 muestra de forma gráfica el funcionamiento del algoritmo de conmutabilidad.

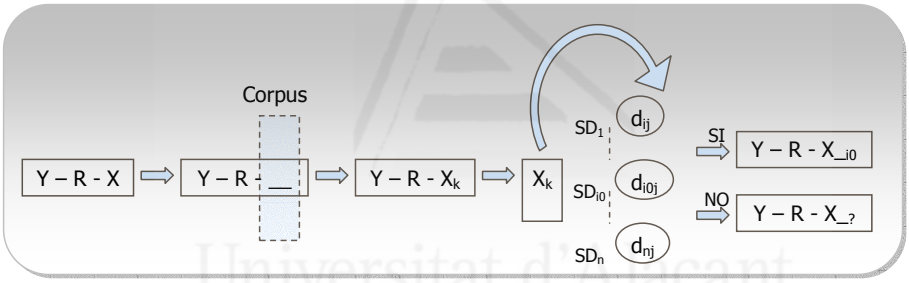


Figura 5.21. Prueba de conmutabilidad

Universitat d'Alacant
Universidad de Alicante

Para un patrón [Y - R - X] cualquiera, se deja libre la posición que ocupa la palabra ambigua. Se buscan en el corpus ocurrencias del patrón con palabras que conmutan en el lugar de la palabra ambigua. Esas palabras conmutables se buscan en los Discriminadores de Sentido para cada sentido de la palabra ambigua. El sentido elegido finalmente es aquel que comparte mayor número de palabras conmutadas en su conjunto de Discriminadores de Sentido.

Una de las ventajas de este algoritmo es que no necesita corpus anotados semánticamente, ya que actúa directamente sobre palabras y no sobre sus sentidos.

La arquitectura del sistema de desambiguación automática es el mostrado en la Figura 5.22.

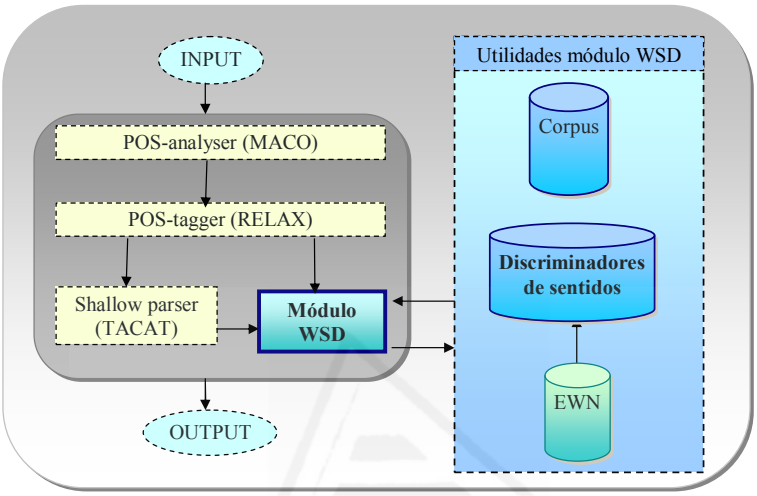


Figura 5.22. Arquitectura sistema

Previamente a la obtención de los patrones sintácticos y a la búsqueda en el corpus de información paradigmática es necesario un preproceso del texto. Se deben determinar los lemas de las palabras y sus relaciones sintácticas. Para ello, se utiliza un analizador morfosintáctico para español (Civit (2003)) y un desambiguador morfológico (Atserias et al. (1998)).

5.3.3 Heurísticas

Además de utilizar la información paradigmática obtenida a partir de la sustitución de la palabra ambigua en el patrón sintáctico, se va a utilizar la información proporcionada por la oración donde aparece la palabra ambigua. Dado que el contexto donde aparece la palabra ambigua (oración) ofrece información

muy valiosa, es interesante utilizar también esta información sobre los conjuntos de Discriminadores de Sentidos.

Se tendrán por tanto, dos fuentes de información:

- **C1:** El conjunto de palabras correspondientes a la información paradigmática obtenida a partir del patrón sintáctico donde aparece la palabra ambigua.
- **C2:** El conjunto de todos los nombres de la oración donde aparece la palabra ambigua.

Sobre cada fuente de información (C1 y C2) se aplica una heurística:

- **H1:** Sobre la información paradigmática.
- **H2:** Sobre la información proporcionada por la oración.

Cada heurística intersecta el conjunto C1 o C2 con los conjuntos de Discriminadores de Sentidos de la palabra ambigua. De esta forma, aquella intersección que dé como resultado un conjunto no vacío de elementos será el sentido elegido. Si cada heurística devuelve un sentido distinto, se conservan ambos. El objetivo de este método es el de filtrar los sentidos inadecuados, sin perder el sentido correcto al tratar de elegir una de las posibles opciones en caso de desacuerdo.

Universitat d'Alacant
Universidad de Alicante

5.3.4 Ejemplo de aplicación

Supongamos que tenemos el siguiente texto del cual queremos obtener el sentido de la palabra *"órgano"*:

"Los enormes y continuados progresos científicos y técnicos de la Medicina actual han logrado hacer descender espectacularmente la mortalidad infantil, erradicar multitud de enfermedades hasta hace poco mortales, sustituir mediante trasplante o implantación de ***órganos*** *dañados o partes del cuerpo inutilizadas y alargar las expectativas de vida."*

El primer paso es detectar el tipo de patrón sintáctico en el que se encuentra la palabra a desambiguar. Para ello se utilizan los patrones sintácticos previamente adquiridos y en el caso de que sea

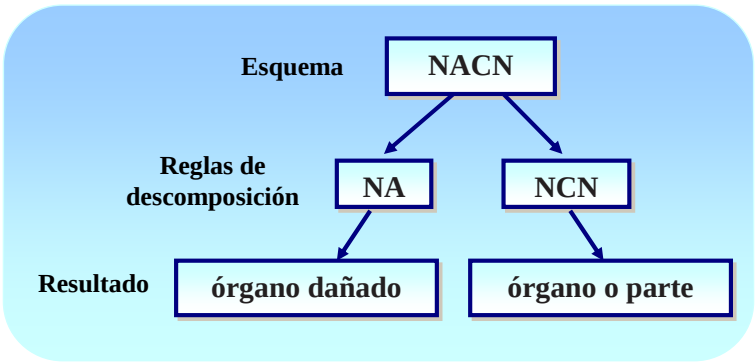


Figura 5.23. Extracción de patrones

un patrón compuesto se utilizan las reglas de descomposición de patrones. La Figura 5.23 muestra la extracción de estos patrones.

El siguiente paso, una vez se han extraído los patrones sintácticos, es obtener información paradigmática presente tanto en corpus como en el contexto que rodea a la palabra ambigua. La Figura 5.24 muestra el resultado de este proceso.

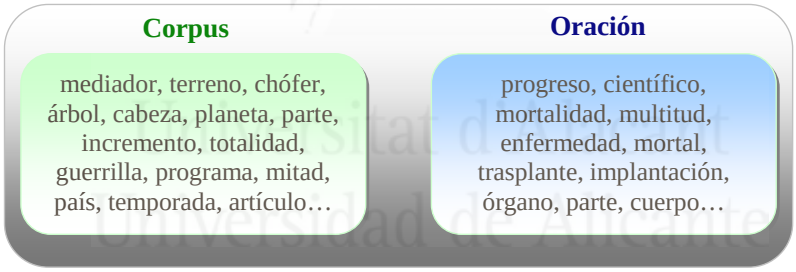


Figura 5.24. Información paradigmática

Una vez extraída la información paradigmática se compara con los conjuntos de Discriminadores de Sentidos asociados a *"órgano"*. Los cinco conjuntos asociados a cada sentido de *"órgano"* se muestran en la Tabla 5.8.

órgano#1: órgano vegetal, espora, flor, pera, manzana, bellota, hinojo, semilla, poro, píleo, carpóforo, ...
órgano#2: agencia, unidad administrativa, banco central, servicio secreto, seguridad social, FBI, ...
órgano#3: parte del cuerpo, trozo, músculo, riñón, oreja, ojo, glándula, lóbulo, tórax, dedo, articulación, rasgo, facción, ...
órgano#4: instrumento de viento, instrumento musical, mecanismo, aparato, teclado, pedal, corneta, ...
órgano#5: periódico, publicación, medio de comunicación, método, serie, serial, número, ejemplar, ...

Tabla 5.8. Discriminadores de Sentidos para "órgano"

Tras comparar el conjunto de palabras del eje paradigmático con cada uno de los conjuntos de Discriminadores de Sentidos, se aplican las dos heurísticas, tal y como muestra la Figura 5.25.

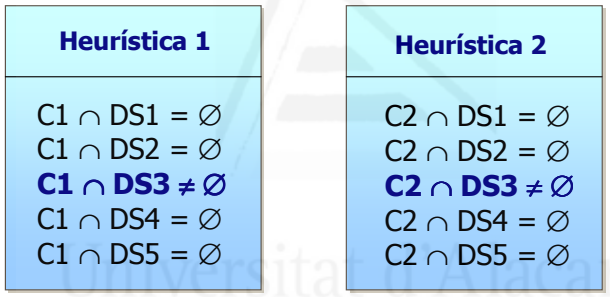


Figura 5.25. Heurísticas

Universitat d'Alacant
Universidad de Alicante

Finalmente el sentido seleccionado es:

órgano#3: *A fully differentiated structural and functional unit in an animal that is specialized for some particular function.*

Capítulo 6

Experimentación y evaluación

En este capítulo se describe todo el proceso de evaluación de los sistemas de WSD implementados en este trabajo, así como su integración y aplicación en otras tareas de PLN. La primera parte del capítulo se centra en el marco de trabajo para la evaluación de sistemas de WSD, presentando las diferentes ediciones de la competición Senseval[1] hasta la actualidad . En la segunda parte, se presenta la evaluación de nuestros sistemas comparando los resultados obtenidos con otros sistemas participantes en Senseval. Finalmente, los sistemas de desambiguación se aplican a otras tareas de PLN que requieren de un módulo de desambiguación semántica para mejorar sus resultados.

Universidad de Alicante

6.1 Competiciones de evaluación

A continuación se describen las tareas que se organizaron en las distintas ediciones de Senseval y algunos de los sistemas que han participado en esta competición en sus diferentes ediciones.

[1] http://www.senseval.org

6.1.1 SENSEVAL: Evaluation Exercises for the Semantic Analysis of Text

SENSEVAL es una competición de evaluación organizada en la línea de otras competiciones como (D)ARPA (HLT1 (1993)), MUC (MUC (1995), MUC (1998)) y TREC (D. (1995), D. (1996)). La primera competición se realizó en el año 1998 con veinticinco sistemas participantes clasificados en dos categorías diferentes: sistemas supervisados y sistemas no supervisados. Los sistemas supervisados que participaron requerían unos datos de entrenamiento anotados semánticamente. Mientras que para los sistemas no supervisados este tipo de información no era necesaria. En esta primera edición se utilizó un lexicon especial para los distintos conjuntos de sentidos: HECTOR (Atkins (1992)), que fue creado por la Universidad de Oxford. A los sistemas participantes se les proporcionaron los datos de entrenamiento, que eran textos anotados con el sentido correcto tomando como referencia HECTOR. La evaluación se llevó a cabo poniendo a prueba los distintos sistemas con una serie de ejemplos sin etiquetar, estos ejemplos debían ser anotados con el sentido correcto de cada palabra. Los idiomas utilizados en esta evaluación fueron: Inglés (18 sistemas), Francés (5 sistemas) e Italiano (2 sistemas).

En esta primera competición los items a desambiguar se restringieron a un conjunto de 40 palabras (tarea "Lexical Sample"). Este conjunto de palabras se estableció de forma aleatoria, de manera que fueron seleccionadas aquellas palabras con suficiente contexto y ejemplos, utilizando la estrategia descrita por (Kilgarriff (1998a)).

En (Kilgarriff y Palmer (2000)) se describe el ámbito de esta competición así como los problemas surgidos en la elección de los corpus y el repositorio de sentidos. Con respecto a la tarea *"English Lexical Sample"* en (Kilgarriff y Rosenzweig (2000)) se realiza un estudio exhaustivo de todos los sistemas propuestos en esta competición. Los sistemas que mejores resultados proporcionaron fueron los sistemas supervisados de la Universidad de Durham (Hawkins y Nettleton (2000)) y de la Universidad de John Hopkins (Yarowsky (2000b)).

Tras esta primera aproximación para evaluar los diferentes sistemas de WSD, se han realizado tres competiciones más: SENSEVAL-2 ACL (2001), SENSEVAL-3 ACL (2004) y la última de ellas SEMEVAL ACL (2007).

6.1.2 SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems

En esta edición de SENSEVAL doce idiomas fueron evaluados y se presentaron alrededor de noventa sistemas. Una de las principales diferencias con la edición anterior fue la adición de una nueva tarea: *"All-Words"*, que requería que los sistemas fueran capaces de etiquetar con el sentido correcto todas las palabras con contenido semántico de los textos proporcionados.

Las tareas propuestas fueron tres (entre paréntesis se muestra el número de sistemas participantes en cada idioma):

- **Tarea All-words.** Checo (1), holandés (datos no disponibles para evaluación), inglés (21) y estonio (2).
- **Tarea Lexical Sample.** Euskera (3), inglés (27), italiano (2), japonés (7), coreano (2), español (12) y sueco (8).
- **Tarea Traducción.** Japonés (9).

La tarea *"Lexical Sample"* tiene como objetivo la obtención del sentido correcto de una única palabra por frase, mientras que en la tarea *"All words"*, los sistemas deben desambiguar semánticamente todas las palabras, a excepción de las funcionales (conjunciones, preposiciones, artículos, etc.), que aparecen en las frases de los corpus proporcionados. La tarea de traducción, en la que sólo se participó con la lengua japonesa, es un subtipo de *"Lexical Sample"* porque sólo hay que desambiguar una única palabra. La diferencia es que el sentido de la palabra se define de acuerdo a su traducción.

El repositorio utilizado para establecer los sentidos de las palabras fue WordNet 1.7 para el idioma inglés y EuroWordNet para el resto de idiomas.

6.1.3 SENSEVAL-3: Evaluation exercises for Word Sense Disambiguation

La tercera competición de SENSEVAL tuvo lugar en Julio de 2004 Barcelona (España). En esta edición se organizaron catorce tareas:

- **Tarea 1.** *English all words.* (64 sistemas)(Snyder y Palmer (2004)) Tal y como se hizo en el Senseval2, en esta nueva edición se etiquetaron aproximadamente 5000 palabras extraídas del corpus de Penn Treebank, tomando como referencia WordNet 1.7.1. Se etiquetaron nombres, adjetivos y adverbios haciéndose dos revisiones para formalizar criterios.
- **Tarea 2.** *Italian all words.* (7 sistemas) (Ulivieri et al. (2004)) En esta edición además de proponer una tarea de "lexical sample" para italiano también se propuso la tarea de "all words". A cada participante se le proporcionó un pequeño conjunto de textos de aproximadamente 5000 palabras extraídos del corpus Italian Treebank. Las palabras etiquetadas fueron nombres, verbos, adjetivos, adverbios y nombres propios, todas ellas atendiendo a la anotación de ItalWordNet (Corazzari y Alonge (2001)).
- **Tarea 3.** *Basque lexical sample.* (8 sistemas) (Agirre et al. (2004)) En esta tarea se evaluaron sistemas supervisados y semi-supervisados para WSD. Cada participante fue provisto de un pequeño conjunto de ejemplos etiquetados y un conjunto más amplio de ejemplos no etiquetados para unas 40 palabras. Todas las palabras fueron etiquetadas con la versión de WordNet 1.6 (aunque se puede obtener fácilmente su equivalente en la versión de WordNet 1.7). Esta tarea fue coordinada junto con otras tareas de lexical sample (Catalán, Inglés, Italiano, Rumano, Español) para tener en común al menos 10 de las palabras utilizadas en la evaluación.
- **Tarea 4.** *Catalan lexical sample.* (8 sistemas) (Màrquez et al. (2004b)) Al igual que en caso del Euskera también se hizo una tarea de lexical sample para Catalán. Compartiendo la etiquetación con WordNet 1.6 y su adaptación a WordNet 1.7. También se proporcionaron textos de entrenamiento etiquetados y

textos no etiquetados para los sistemas supervisados y semi-supervisados que participaron.

- **Tarea 5.** *Chinese lexical sample.* (16 sistemas) En esta tarea se utilizaron tres tipos de datos: diccionario, datos de entrenamiento y datos de evaluación. El diccionario contenía entradas para 20 palabras distintas. Para cada palabra habían definidos varios sentidos basados en el recurso HowNet (Gan y Wong (2000)). Para cada sentido, la entrada del diccionario listaba: un identificador para el sentido, la categoría sintáctica de la palabra, una definición y una traducción al inglés, así como alguna información adicional. Los datos de entrenamiento consistían en 20-100 ejemplos por palabra, con más ejemplos para aquellas palabras con un número más elevado de sentidos. Dos conjuntos de entrenamiento fueron distribuidos: uno con etiquetación sintáctica y otro sin esa información.
- **Tarea 6.** *English lexical sample.* (65 sistemas) (Mihalcea et al. (2004a))
  Los datos en esta tarea fueron obtenidos a partir de la interfaz del Open Mind Word Expert (OMWE) (Chklovski y Mihalcea (2002)). Para asegurar la fiabilidad se extrajeron dos etiquetas por item y se realizaron diversos tests con la finalidad de llegar a un acuerdo entre las distintas anotaciones de los etiquetadores. La elección de OMWE como interfaz para obtener los datos de la tarea fue debida a su probada calidad en otras evaluaciones. Se extrajeron alrededor de 60 palabras ambiguas entre nombres, adjetivos y verbos. Parte del test de evaluación fue creado por el Departamento de Lingüística de la Universidad del Norte de Texas (UNT). Otra parte del test de evaluación fue extraído a partir de corpus etiquetado de la web. Se utilizó la versión de WordNet 1.7.1 para nombres y adjetivos y Wordsmyth [2] para verbos. Además también se distribuyeron los agrupamientos de sentidos que posibilitaban una evaluación menos restrictiva (coarse) frente a una evaluación más estricta (fine). Además también se distribuyó el mapeo entre la anotación de Wordsmyth y WordNet.

[2] `http://www.wordsmyth.net/`

- **Tarea 7.** *Italian lexical sample.* (11 sistemas) (Magnini et al. (2004)) En esta tarea sigue los mismos principios que las tareas de lexical sample para Euskera y Catalán mencionadas anteriormente. Se utilizó para la anotación de sentidos el MultiWordNet Italiano (Pianta et al. (2002)), que fue especialmente desarrollado para la tarea.
- **Tarea 8.** *Romanian lexical sample.* (8 sistemas) (Mihalcea et al. (2004b))
  En esta tarea se seleccionaron 50 palabras, cubriendo nombres, adjetivos, verbos y adverbios, con distintos grados de ambigüedad. Para cada palabra se extrajeron un conjunto de ejemplos a partir de un extenso corpus en Rumano. Los sentidos y las expresiones de palabras múltiples fueron extraídos del WordNet Rumano o de DEX [3] un reconocido diccionario del Rumano. Los datos fueron extraídos a partir de la interfaz del OMWE, edición en rumano.
- **Tarea 9.** *Spanish lexical sample.* (18 sistemas) (Màrquez et al. (2004a))
  Al igual que para Catalán, Euskera, Inglés e Italiano también se organizó la tarea lexical sample para Español. El repertorio de sentidos fue obtenido a partir de WordNet 1.6 con su respectivo mapeo a WordNet 1.7. Esta tarea se coordinó con las demás tareas de lexical sample para compartir al menos 10 de las palabras propuestas para desambiguar.
- **Tarea 10.** *Automatic subcategorization acquisition.* (35 sistemas) (Preiss y Korhonen (2004))
  En esta tarea se evaluaron diversos sistemas de WSD en el contexto de subcategorización automática. La tarea se restringió a 30 verbos, de elevada frecuencia de aparición y con múltiples sentidos. En este caso se publicaron los 30 verbos de la tarea pero no se proporcionó ningún corpus de entrenamiento. El corpus de evaluación consistió en alrededor de 1000 instancias para cada verbo, que se debían anotar con el sentido correcto, de acuerdo a la versión de WordNet 1.7.1.

[3] `http://dexonline.ro/`

- **Tarea 11.** *Multilingual lexical sample.* (23 sistemas) (Chklovski et al. (2004))
  El objetivo de esta tarea era crear un marco de referencia para la evaluación de sistemas de Traducción Automática, centrándose en la traducción de palabras ambiguas. Esta tarea era muy similar a la tarea lexical sample, excepto que en lugar de utilizar el inventario de sentidos de un diccionario se utilizó la propuesta de Resnik y Yarowsky usando las traducciones de las palabras en otra lengua como "inventario". Los textos originales eran en Inglés y la anotación para las palabras se debía hacer la traducción en otro idioma. La tarea se restringió a textos Inglés-Francés e Inglés-Hindi con alrededor de 50 palabras ambiguas por pareja de idiomas.
- **Tarea 12.** *WSD of WordNet glosses.* (36 sistemas)(Litkowski (2004b)) Relacionado con WordNet se desarrolló eXtended WordNet (Harabagiu et al. (1999)), un recurso que enriquece la versión inicial de WordNet añadiendo la contenido morfológico y semántico, a los términos de las glosas de WordNet. El proceso para obtener una buena anotación es costoso y en muchos casos se realiza de forma manual. El objetivo de esta tarea era desarrollar un método de anotación automática tomando como corpus de evaluación las glosas previamente etiquetadas en eXtended WordNet. En el marco de la tarea *"all-words"*, se debían etiquetar nombres, verbos, adverbios y adjetivos, con la salvedad de que ningún contexto era proporcionado. Los sistemas participantes podían utilizar cualquier información adicional como synsets, la jerarquía de WordNet y otro tipo de relaciones en WordNet.
- **Tarea 13.** *Semantic Roles.* (36 sistemas) (Litkowski (2004a)) Utilizando como base una porción del corpus anotado de FrameNet, los sistemas debían realizar la anotación de roles semánticos siguiendo las métricas del estudio de Gildea y Jurafsky (Gildea y Jurafsky (2002)).
- **Tarea 14.** *Logic Forms.* (26 sistemas) (Rus (2004)) El objetivo de esta tarea era transformar oraciones formuladas en inglés en su correspondiente notación de lógica de primer orden. Cada

palabra con contenido semántico se correspondía con un predicado.

6.1.4 SENSEVAL-4/SEMEVAL-1: 4th International Workshop on Semantic Evaluations

La última edición SENSEVAL [4] tuvo lugar en Junio de 2007 en Praga (República Checa). En esta edición se organizaron dieciocho tareas (la tarea 3 se canceló):

- **Tarea 1.** Evaluating WSD on Cross-Language Information Retrieval. (2 sistemas) (Agirre et al. (2007))
- **Tarea 2.** Evaluating Word Sense Induction and Discrimination Systems. (6 sistemas) (Agirre y Soroa (2007))
- **Tarea 4.** Classification of Semantic Relations between Nominals. (15 sistemas) (Girju et al. (2007))
- **Tarea 5.** Multilingual Chinese-English Lexical Sample. (6 sistemas) (Jin et al. (2007))
- **Tarea 6.** Word Sense Disambiguation of Prepositions. (5 sistemas) (Litkowski y Hargraves (2007))
- **Tarea 7.** Coarse Grained English All Words Task. (12 sistemas) (Navigli et al. (2007))
- **Tarea 8.** Metonymy Resolution at SemEval 2007. (5 sistemas) (Markert y Nissim (2007))
- **Tarea 9.** Multilevel Semantic Annotation of Catalan and Spanish. (2 sistemas) (Màrquez et al. (2007))
- **Tarea 10.** English Lexical Substitution Task. (McCarthy y Navigli (2007))
- **Tarea 11.** English Lexical Sample Task via English-Chinese Parallel Text. (Ng y Chan (2007))
- **Tarea 12.** Turkish Lexical Sample Task. (Orhan et al. (2007))
- **Tarea 13.** Web People Search. (Artiles et al. (2007))
- **Tarea 14.** Affective Text. (Strapparava y Mihalcea (2007))
- **Tarea 15.** TempEval Temporal Relation Identification. (Verhagen et al. (2007))

[4] http://nlp.cs.swarthmore.edu/semeval/

- **Tarea 16.** Evaluation of Wide Coverage Knowledge Resources. (Cuadros y Rigau (2007))
- **Tarea 17.** English Lexical Sample, SRL and All Words. (Pradhan et al. (2007))
- **Tarea 18.** Arabic Semantic Labeling. (Diab et al. (2007))
- **Tarea 19.** Frame Semantic Structure Extraction. (Baker et al. (2007))

6.2 Participación en Senseval

Para la evaluación de los diferentes sistemas de WSD, es necesario determinar un inventario de sentidos con el que anotar las palabras polisémicas de los textos. Además, se debe definir un conjunto de corpus sobre los que realizar la anotación y posterior evaluación de los resultados. Junto con el establecimiento de un repositorio de sentidos, se plantean otro tipo de problemas, como por ejemplo, el cómo decidir el grado de distinción entre un sentido u otro. Este problema se denomina nivel de granularidad de sentidos (Edmonds y Kilgarriff (1998)), donde se pueden distinguir dos tipos: granularidad fina (*"fine-grained"*) y granularidad gruesa (*"coarse-grained"*). Los repositorios de sentidos con granularidad fina tienen como característica poseer una división muy detallada de los sentidos pero con un alto nivel de ambigüedad. Sin embargo, los repositorios de sentidos con granularidad gruesa proporcionan una división muy general de los sentidos con un nivel muy bajo de ambigüedad.

Para las diferentes evaluaciones dentro de la competición SENSEVAL, se ha utilizado como repositorio de sentidos WordNet, caracterizado por tener una granularidad fina, donde existe una división de los sentidos muy detallada. Todos los sistemas participantes deben anotar una serie de palabras polisémicas de acuerdo a los diferentes sentidos de WordNet. Finalmente, la evaluación de la efectividad de cada sistema se realiza estableciendo una comparación con respecto a una anotación manual de los corpus de evaluación.

Las medidas utilizadas para la evaluación de los sistemas de WSD son las siguientes:

$$Precision = \frac{Instancias\ \ correctas\ \ contestadas}{Instancias\ \ contestadas} \tag{6.1}$$

$$Cobertura = \frac{Instancias\ \ correctas\ \ contestadas}{Total\ \ instancias} \tag{6.2}$$

A continuación se describen las tareas en las que nuestros sistemas de WSD han participado, junto con los resultados obtenidos y su clasificación respecto a otros sistemas participantes.

6.2.1 DRelevant: All Words

La evaluación del método de desambiguación DRelevant presentado en este trabajo, se ha realizado sobre los textos de la tarea *"English all-words"* de SENSEVAL-2. Esta tarea consiste en desambiguar todas las palabras con contenido semántico (nombres, verbos, adjetivos y adverbios), que aparecen en los textos proporcionados, para su posterior evaluación. El número total de palabras a desambiguar (nombres, verbos, adjetivos y adverbios) supone un total de 2473 instancias.

Para evaluar la eficiencia del método, se han adoptado diferentes criterios atendiendo al tamaño del contexto seleccionado y al número de dominios empleado para realizar el proceso de desambiguación. En las siguientes secciones se detallan cada una de los experimentos realizados.

6.2.1.1 Experimento 1: Oración como contexto.

En este primer experimento, el contexto seleccionado para realizar el proceso de desambiguación es la oración. Es decir, a partir del corpus proporcionado se extraen todas las oraciones, se analizan morfológicamente mediante el *"Tree-tagger"* y se guardan los nombres, verbos, adjetivos y adverbios de cada oración en diferentes ficheros, que serán la entrada para el sistema DRelevant.

En la Tabla 6.1, se muestra la eficiencia del método con el criterio de utilizar como contexto la oración.

Contexto: Oración		
Precisión	**Cobertura**	**Cobertura Absoluta**
44 %	32 %	73 %

Tabla 6.1. Medida de la eficiencia utilizando como contexto la oración

Tras el proceso de desambiguación, el número de palabras correctamente desambiguadas no superó el 50 % del número total de instancias. Estos resultados son debidos en gran medida a la escasa información que aporta el contexto de la oración, ya que, el número de palabras proporcionadas por la oración no es suficiente para obtener una buena información contextual y determinar correctamente los sentidos.

6.2.1.2 Experimento 2: Ventana de 100 palabras como contexto.

En esta segunda prueba, se establece como contexto una ventana de 100 palabras alrededor de cada palabra ambigua. Es decir, para cada una de las palabras a desambiguar se extraen las 100 palabras con contenido semántico que la rodean (50 palabras previas y 50 palabras posteriores). Por ejemplo, si quisiéramos desambiguar la palabra *'sound'*, el contexto extraído sería el siguiente:

"(50 palabras previas)... Then, at a signal, the ringers begin varying the order in which the bells ***sound*** *without altering the steady rhythm of the striking...(50 palabras posteriores)"*

En caso de que no se pudieran extraer las 50 palabras anteriores a la palabra a desambiguar, porque se ha llegado al comienzo del texto, el contexto de 100 palabras se completaría con la información de las palabras posteriores a la palabra ambigua. Del mismo modo, en caso de que no se pudieran extraer las 50 palabras posteriores a la palabra a desambiguar porque se ha llegado

al final del texto, el resto, se extraería a partir de las palabras anteriores a la palabra ambigua.

En la Tabla 6.2 se muestra la eficiencia del método al utilizar como contexto una ventana de 100 palabras. En este caso tanto la precisión, como la cobertura y la cobertura absoluta sufren un incremento, llegando a alcanzar el 47 %, el 38 % y el 81 % respectivamente.

Contexto: Ventana de 100 palabras		
Precisión	**Cobertura**	**Cobertura Absoluta**
47 %	38 %	81 %

Tabla 6.2. Medida de la eficiencia utilizando como contexto una ventana de 100 palabras

A la vista de los resultados obtenidos, es evidente la necesidad de disponer de un contexto lo suficientemente amplio, para establecer correctamente los sentidos de las palabras.

6.2.1.3 Experimento 3: Reducción y agrupación de los dominios.

En este experimento, se intenta minimizar el nivel de especialización de los dominios. Este proceso se ha realizado partiendo de la estructuración jerárquica de los dominios en WordNet Domains. En este caso, se agrupan dentro de un dominio que se encuentra en un nivel superior de la jerarquía, aquellos subdominios que dependen de él. Es decir, se agrupan dentro de un mismo dominio el conjunto de dominios que pertenecen a su misma jerarquía pero que están en los niveles inferiores. Por ejemplo, en el caso de la jerarquía del dominio MEDICINE, se encuentran por debajo de él los dominios: DENTISTRY, PHARMACY, PSYCHIATRY, RADIOLOGY y SURGERY. Estos dominios se engloban dentro del dominio MEDICINE, y así se reduce el espacio de búsqueda y el grado de especialización. Obteniendo finalmente 43 dominios, sobre los 165 dominios iniciales.

La reducción del nivel de especialización de los dominios requiere una nueva anotación de los sentidos de las palabras de WordNet y por tanto, una nueva obtención de los dominios relevantes junto con su correspondiente ratio de asociación. Esta tarea se ha realizado previamente a la aplicación del sistema DRelevant.

Dado que los resultados a partir de una ventana contextual de 100 palabras han demostrado obtener mejores resultados en el proceso de anotación de sentidos, este experimento, se realizó manteniendo la ventana de 100 palabras y utilizando la reducción del nivel de especialización de los dominios.

En este caso, tras analizar los resultados obtenidos, se vuelve a incrementar el número de palabras correctamente desambiguadas, y por consiguiente, la precisión también sufre un incremento, tal y como se muestra en la Tabla 6.3.

Reducción del número de dominios		
Precisión	**Cobertura**	**Cobertura Absoluta**
48 %	41 %	85 %

Tabla 6.3. Medida de la eficiencia reduciendo el nivel de especialización de los dominios

Se hace patente, por tanto, que la reducción del nivel de especialización de los dominios, influye positivamente en el proceso de desambiguación.

Universidad de Alicante

6.2.1.4 Experimento 4: Desambiguación a nivel de dominio.

El último experimento realizado, se basa en la necesidad de reducir el número de sentidos de una misma palabra proporcionados por WordNet. Es decir, en WordNet, la distinción entre los distintos sentidos de una palabra es en algunos casos muy difícil de establecer, es lo que se denomina granularidad fina, como se comentaba al principio de esta sección. Para intentar reducir esta granularidad se agrupan aquellos sentidos etiquetados con el

mismo dominio. De esta forma, el resultado de la desambiguación para una palabra, no sería un único sentido, sino todos aquellos sentidos de la palabra que tengan asociado el mismo dominio que se obtenga tras el proceso de desambiguación. Por ejemplo, supongamos que tras el proceso de desambiguación de la palabra *"bank"*, se obtiene el dominio ECONOMY, entonces daríamos como resultado los tres sentidos asociados a este dominio: *bank#1*, *bank#3* y *bank#6*.

La reducción de la granularidad de WordNet se realizó utilizando los 165 dominios de la jerarquía de WordNet Domains. La opción de reducir la granularidad a partir de reducir el nivel de especialización de dominios no se ha planteado, porque se reducen dramáticamente el número de sentidos de las palabras.

En este caso, los resultados obtenidos reportan una precisión del 54 %, tal y como se muestra en la Tabla 6.4.

Desambiguación a nivel de dominio		
Precisión	**Cobertura**	**Cobertura Absoluta**
54 %	43 %	80 %

Tabla 6.4. Medida de la eficiencia desambiguando a nivel de dominio

Así pues, la agrupación de sentidos como cabía esperar obtiene mejores resultados. En muchos casos, la distinción de sentidos es demasiado fina, y es muy difícil establecer la línea que diferencia un sentido de otro, por tanto, la agrupación resuelve este dilema.

6.2.1.5 Comparativa con otros sistemas.

La evaluación del sistema DRelevant aquí presentado sobre la tarea *"English all-words"* de SENSEVAL-2, tiene como finalidad el poder establecer una comparativa con el resto de sistemas no supervisados, que participaron en esta edición. Esta comparación se realiza atendiendo a las medidas de precisión y cobertura.

En la Tabla 6.5 se muestran los resultados obtenidos por los distintos sistemas que participaron en la tarea *"English all-words"* de SENSEVAL-2.

	Sistema	Precision	Cobertura
1	SMWaw-	0.69	0.690
2	Ave-Antwerp	0.636	0.636
3	LIA-Sinequa-AllWords	0.618	0.618
4	David-fa-UNED-AW-T	0.575	0.569
5	David-fa-UNED-AW-U	0.556	0.550
6	Gchao2-	0.475	0.454
7	Gchao3-	0.474	0.453
8	Ken-Litkowski-clr-aw	0.451	0.451
9	Gchao-	0.500	0.449
Exp4	DRelevant-4	0.54	0.43
Exp3	DRelevant-3	0.48	0.41
Exp2	DRelevant-2	0.47	0.38
10	cm.guo-usm-english-tagger2	0.360	0.360
11	Magnini2-irst-eng-all	0.748	0.357
12	Cmguo-usm-english-tagger	0.345	0.338
13	c.guo-usm-english-tagger3	0.336	0.336
Exp1	DRelevant-1	0.44	0.32
14	Agirre2-ehu-dlist-all	0.572	0.291
15	Judita-	0.440	0.200
16	Dianam-system3ospdana	0.545	0.169
17	Dianam-system2ospd	0.566	0.169
18	Dianam-system1	0.598	0.140
19	Woody-IIT2	0.328	0.038
20	Woody-IIT3	0.294	0.034
21	Woody-IIT1	0.287	0.033

Tabla 6.5. Comparación de los resultados de los distintos sistemas participantes en la tarea *"English all-words"* de SENSEVAL-2.

A continuación vamos a evaluar la posición alcanzada por nuestro sistema en cada uno de los experimentos realizados:

- **Experimento 1.** Este experimento trata de evaluar la eficiencia del método de desambiguación propuesto, utilizando como contexto las palabras de la oración donde aparece la instancia a desambiguar. En este caso, la precisión obtenida es de

un 44 % y la cobertura es de un 32 %. Con estos resultados nuestro sistema se situaría en la posición 14 por delante del sistema de Agirre2-ehu-dlist-all.

- **Experimento 2.** En este experimento el contexto utilizado se amplía mediante la utilización de una ventana de 100 palabras alrededor de la instancia a desambiguar. En este caso, la precisión obtenida es de un 47 % y la cobertura es de un 38 %. Con los resultados obtenidos en este experimento nuestro sistema se situaría en la posición 10 por delante del método de Magnini2-irst-eng-all. Este dato supone que nuestro método mejora los resultados obtenidos por el método de Magnini, e indica que la utilización de las glosas de WordNet Domains para extraer el recurso léxico Dominios Relevantes y su posterior utilización en un método de WSD ofrece buenos resultados.
- **Experimento 3.** En este experimento se reduce el nivel de especialización de los dominios y se utiliza una ventana de 100 palabras alrededor de la instancia a desambiguar. Se obtiene un 48 % de precisión y un 41 % de cobertura. Los resultados obtenidos en este experimento mejoran la precisión y la cobertura del experimento anterior, pero no suponen un cambio en la posición alcanzada por nuestro sistema con respecto a los otros participantes de Senseval-2
- **Experimento 4.** En este experimento se trata de reducir el problema de la granularidad fina de WordNet agrupando aquellos sentidos de una misma palabra, que comparten un mismo dominio. En este caso, la precisión obtenida es de un 54 % y la cobertura es de un 43 %. Aquí ocurre lo mismo que en experimento anterior, los valores de precisión y cobertura sufren una mejora, pero no lo suficiente como para mejorar la posición de nuestro sistema.

En definitiva, la utilización del recurso léxico Dominios Relevantes y su aplicación en un método de WSD, ofrece unos resultados prometedores con respecto a los actuales sistemas de desambiguación automática del sentido de las palabras. Además, el recurso Dominios Relevantes, ofrece información muy útil para relacionar semánticamente diferentes palabras, y puede ser utilizado

como recurso para otras tareas de PLN, tal y como se demuestra en las siguientes secciones: reconocimiento de la implicación textual, discriminación de nombres, etc.

6.2.2 DRelevant mejorado con Extended WordNet

El recurso léxico Dominios Relevantes, utilizado como base de conocimiento del método de desambiguación DRelevant, puede ser mejorado si se utiliza para su construcción Extended WordNet en lugar de WordNet Domains. Como ya se comentó en el capítulo anterior, Extended WordNet proporciona información adicional acerca de los sentidos de las palabras de las glosas. Esta información ha sido utilizada para obtener de nuevo el recurso Dominios Relevantes mejorado.

Utilizando esta nueva versión de DR se han realizado los mismos experimentos para la tarea all-words obteniendo los resultados mostrados en la Tabla 6.6.

Opción	Precisión	Recall
Contexto oración	0.56	0.46
Ventana de 100 palabras	0.61	0.47
Reducción del num. de dominios	0.62	0.50
Desamb. a nivel de dominio	0.63	0.56

Tabla 6.6. Evaluación de WSD DRelevant usando Extended WordNet

Universitat d'Alacant
Universidad de Alicante

A la vista de los resultados, se aprecia una mejora sobre los anteriores experimentos. Se produce un aumento del 12 % a nivel de precisión en el experimento que toma como contexto la oración. Un 4 % de mejora con respecto al experimento que utiliza una ventana de 100 palabras y el experimento que reduce el número de dominios, agrupando aquellos que descienden del mismo concepto. Lo mismo ocurre con el último experimento, que trata de evitar la granularidad fina de WordNet, desambiguando a nivel de dominio en lugar de a nivel de sentido.

Podemos concluir que la elección de los contextos y la información proporcionada por éstos, son fundamentales para la ob-

tención del recurso Dominios Relevantes. Además, una buena calidad de los DR supone una mejora de los resultados del sistema WSD Relevant. De esta forma, con los nuevos resultados nuestro sistema escala cinco posiciones con respecto al resto de sistemas participantes en la tarea *"all-words"*.

6.2.3 R2D2: English All Words y English Lexical Sample

Dentro del marco del proyecto R2D2 (Recuperación de Respuestas en Documentos Digitalizados) [5] se participó en la tercera edición de Senseval. El objetivo de esta participación fue evaluar el resultado de la combinación de diferentes sistemas de WSD dentro de las tareas English All-Words y English Lexical Sample.

Para llevar a cabo el experimento se combinaron tanto sistemas supervisados como sistemas no supervisados. De forma que se intentó paliar la falta en muchos casos de ejemplos de entrenamiento para los sistemas supervisados que impedía la correcta detección de los sentidos, incorporando métodos no supervisados que no necesitaban ejemplos de entrenamiento.

Los sistemas participantes para la tarea English All-words fueron cuatro: Maximum Entropy, UPV-SHMM-AW, DRelevant y CIAOSENSO. En la tarea English Lexical Sample participaron también cuatro sistemas: DRelevant, CIAOSENSO, LVQ-JAEN-ELS y Maximum Entropy. La Tabla 6.7 muestra las características de cada uno de los sistemas participantes.

6.2.3.1 R2D2: English All Words.

En la Tabla 6.8 se presentan los resultados obtenidos para los diferentes sistemas participantes, tanto supervisados como no supervisados. En este caso, se han tomado como válidas aquellas respuestas anotadas como desconocidas por todos los sistemas. De esta forma, tanto precisión como cobertura coinciden debido a que siempre se contesta el 100 % de las instancias. El sistema R2D2 resultado de la combinación de sistemas WSD supervisados

[5] Proyecto financiado por el Ministerio de Ciencia y Tecnología. TIC2003-07158-C04-01

Sistemas	Descripción
Maximum entropy	Sistema supervisado basado en los modelos de probabilidad de Máxima entropía. Este sistema utiliza un conjunto de características y un conjunto de ejemplos de entrenamiento extraídos del corpus Semcor para resolver la ambigüedad (Suárez y Palomar (2002)).
UPV-SHMM-AW	Sistema supervisado basado en modelos especializados ocultos de Markov (Specialized Hidden Markov Models) (Molina et al. (2002)).
DRelevant	Sistema no supervisado basado en la adquisición de conocimiento a través de WordNet Domains (Montoyo et al. (2003)).
CIAOSENSO	Sistema no supervisado basado en densidad conceptual, frecuencia de sentidos de WordNet y WordNet Domains (Rosso et al. (2003)).
LVQ-Jaen-ELS	Sistema supervisado basado en redes neuronales utilizando LVQ, integrando Semcor y varias relaciones semánticas de WordNet (Vega et al. (2003)).

Tabla 6.7. Sistemas participantes en el equipo R2D2

y no supervisados, se coloca en cuarto lugar respecto al resto de sistemas participantes.

El sistema de votación utilizado en esta tarea combina los resultados de los diferentes sistemas de WSD según se muestra en la Figura 6.1.

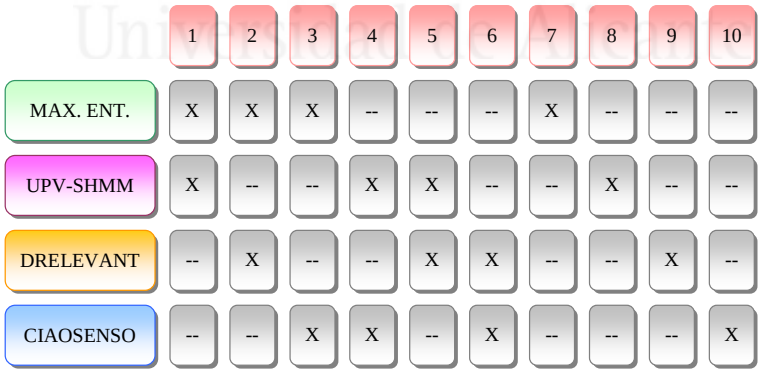


Figura 6.1. Sistema de votación R2D2 All Words

Inicialmente se le da preferencia a aquellos sentidos dados como respuesta por la mayoría de sistemas. Pero en caso de no existir acuerdo, se decide el sentido correcto en un máximo de 10 pasos. Tal y como muestra la Figura 6.1 el primer paso da preferencia a los sistemas supervisados, si no existe acuerdo, el segundo paso comprueba el acuerdo entre Max. Ent. y DRelevant... El sistema de votación continúa hasta que en algún paso exista acuerdo, si no es así, los sentidos que han permanecido sin anotar se deciden en última instancia por un solo sistema, dando siempre preferencia a los sistemas supervisados.

System	Precision/Recall
GAMBL-AW-S	0.652
SenseLearner-S	0.646
Koc University-S	0.641
R2D2: English-all-words	0.626
Meaning-allwords-S	0.624
Meaning-simple-S	0.610
upv-shmm-eaw-S	0.609
LCCaw	0.607
UJAEN-S	0.590
IRST-DDD-00-U	0.583
University of Sussex-Prob5	0.572
University of Sussex-Prob4	0.554
University of Sussex-Prob3	0.551
DFA-Unsup-AW-U	0.548
IRST-DDD-LSI-U	0.501
KUNLP-Eng-All-U	0.500
upv-unige-CIAOSENSO-eaw-U	0.481
merl.system3	0.458
upv-unige-CIAOSENSO2-eaw-U	0.452
merl.system1	0.450
IRST-DDD-09-U	0.446
autoPS-U	0.436
clr04-aw	0.434
merl.system2	0.359
autoPSNVs-U	0.359

Tabla 6.8. Resultados para AllWords con validación de respuestas no anotadas

Universitat d'Alacant
Universidad de Alicante

En la Tabla 6.9 se muestran los resultados obtenidos, ignorando las respuestas no anotadas. De esta forma, ya no existe un 100 % en el número de respuestas anotadas, sino que éste se reduce al no tenerse en cuenta instancias no anotadas con un sentido determinado. De esta forma, la precisión y cobertura difieren en su valor. En este caso, respecto al sistema R2D2, los resultados no difieren en absoluto de los anteriores, debido a que siempre se responden el 100 % de las instancias, utilizando un método de votación entre los diferentes sistemas.

System	Precision	Recall
GAMBL-AW-S	0.651	0.651
SenseLearner-S	0.651	0.642
Koc University-S	0.648	0.639
R2D2: English-all-words	0.626	0.626
Meaning-allwords-S	0.625	0.623
Meaning-simple-S	0.611	0.610
LCCaw	0.614	0.606
upv-shmm-eaw-S	0.616	0.605
UJAEN-S	0.601	0.588
IRST-DDD-00-U	0.583	0.582
University of Sussex-Prob5	0.585	0.568
University of Sussex-Prob4	0.575	0.550
University of Sussex-Prob3	0.573	0.547
DFA-Unsup-AW-U	0.557	0.546
KUNLP-Eng-All-U	0.510	0.496
IRST-DDD-LSI-U	0.661	0.496
upv-unige-CIAOSENSO-eaw-U	0.581	0.480
merl.system3	0.467	0.456
upv-unige-CIAOSENSO2-eaw-U	0.608	0.451
merl.system1	0.459	0.447
IRST-DDD-09-U	0.729	0.441
autoPS-U	0.490	0.433
clr04-aw	0.506	0.431
autoPSNVs-U	0.563	0.354
merl. system2	0.480	0.352

Tabla 6.9. Resultados para AllWords sin validación de respuestas no anotadas

6.2.3.2 R2D2: English Lexical Sample.

En la tarea English Lexical Sample el objetivo es desambiguar una serie de palabras etiquetadas dentro de un corpus: nombres, verbos y adjetivos. El método utilizado para seleccionar el sentido final de cada palabra, es un sistema de votación, donde el sentido más votado por todos los sistemas es el seleccionado. En caso de no existir acuerdo entre varios sentidos se le da prioridad a los sistemas supervisados debido a que demuestran una mejor precisión en esta tarea (sistema similar al presentado en la Figura 6.1). La Tabla 6.2.3.2 muestra los resultados obtenidos en esta tarea.

Tabla 6.10: Sistemas participantes en la tarea English Lexical Sample de Senseval-3

		Fine		Coarse	
System/Team	Description	P	R	P	R
htsa3 U.Bucharest (Grozea)	A Naive Bayes system, with correction of the a-priori frequencies, by dividing the output confidence of the senses by frequency$^{\alpha}$ ($\alpha = 0.2$)	72.9	72.9	79.3	79.3
IRST-Kernels ITC-IRST (Strapparava)	Kernel methods for pattern abstraction, paradigmatic and syntagmatic info.and unsupervised term proximity (LSA) on BNC, in an SVM classifier.	72.6	72.6	79.5	79.5
nusels Nat.U. Singapore (Lee)	A combination of knowledge sources (part-of-speech of neighbouring words, words in context, local collocations, syntactic relations), in an SVM classifier.	72.4	72.4	78.8	78.8
htsa4	Similar to htsa3, with different correction function of a-priori frequencies.	72.4	72.4	78.8	78.8
BCU comb Basque Country U. (Agirre & Martinez)	An ensemble of decision lists, SVM, and vectorial similarity, improved with a variety of smoothing techniques. The features consist of local collocations, syntactic dependencies, bag-of-words, domain features.	72.3	72.3	78.9	78.9
htsa1	Similar to htsa3, but with smaller number of features.	72.2	72.2	78.7	78.7

(continúa ...)

Tabla 6.10: Sistemas participantes en la tarea English Lexical Sample de Senseval-3 (continuación)

		Fine		Coarse	
System/Team	Description	P	R	P	R
rlsc-comb U.Bucharest (Popescu)	A regularized least-square classification (RLSC), using local and topical features, with a term weighting scheme.	72.2	72.2	78.4	78.4
htsa2	Similar to htsa4, but with smaller number of features.	72.1	72.1	78.6	78.6
BCU english	Similar to BCU comb, but with a vectorial space model learning.	72.0	72.0	79.1	79.1
rlsc-lin	Similar to rlsc-comb, with a linear kernel, and a binary weighting scheme.	71.8	71.8	78.4	78.4
HLTC HKUST all HKUST (Carpuat)	A voted classifier combining a new kernel PCA method, a Maximum Entropy model, and a boosting-based model, using syntactic and collocational features	71.4	71.4	78.6	78.6
TALP U.P.Catalunya (Escudero et al.)	A system with per-word feature selection, using a rich feature set. For learning, it uses SVM, and combines two binarization procedures: one vs. all, and constraint learning.	71.3	71.3	78.2	78.2
MC-WSD Brown U. (Ciaramita & Johnson)	A multiclass averaged perceptron classifier with two components: one trained on the data provided, the other trained on this data, and on WordNet glosses. Features consist of local and syntactic features.	71.1	71.1	78.1	78.1
HLTC HKUST all2	Similar to HLTC HKUST all, also adds a Naive Bayes classifier.	70.9	70.9	78.1	78.1
NRC-Fine NRC (Turney)	Syntactic and semantic features, using POS tags and pointwise mutual information on a terabyte corpus. Five basic classifiers are combined with voting.	69.4	69.4	75.9	75.9
HLTC HKUST me	Similar to HLTC HKUST all, only with a maximum entropy classifier.	69.3	69.3	76.4	76.4
NRC-Fine2	Similar to NRC-Fine, with a different threshold for dropping features	69.1	69.1	75.6	75.6

(continúa ...)

Tabla 6.10: Sistemas participantes en la tarea English Lexical Sample de Senseval-3 (continuación)

		Fine		Coarse	
System/Team	Description	P	R	P	R
GAMBL U. Antwerp (Decadt)	A cascaded memory-based classifier, using two classifiers based on global and local features, with a genetic algorithm for parameter optimization.	67.4	67.4	74.0	74.0
SinequaLex Sinequa Labs (Crestan)	Semantic classification trees, built on short contexts and document semantics, plus a decision system based on information retrieval techniques.	67.2	67.2	74.2	74.2
CLaC1 Concordia U. (Lamjiri)	A Naive Bayes approach using a context window around the target word, which is dynamically adjusted	67.2	67.2	75.1	75.1
SinequaLex2	A cumulative method based on scores of surrounding words.	66.8	66.8	73.6	73.6
UMD SST4 U. Maryland (Cabezas)	Supervised learning using Support Vector Machines, using local and wide context features, and also grammatical and expanded contexts.	66.0	66.0	73.7	73.7
Prob1 Cambridge U. (Preiss)	A probabilistic modular WSD system, with individual modules based on separate known approaches to WSD (26 different modules)	65.1	65.1	71.6	71.6
SyntaLex-3 U.Toronto (Mohammad)	A supervised system that uses local part of speech features and bigrams, in an ensemble classifier using bagged decision trees.	64.6	64.6	72.0	72.0
UNED UNED (Artiles)	A similarity-based system, relying on the co-occurrence of nouns and adjectives in the test and training examples.	64.1	64.1	72.0	72.0
SyntaLex-4	Similar to SyntaLex-3, but with unified decision trees.	63.3	63.3	71.1	71.1
CLaC2	Syntactic and semantic (WordNet hypernyms) information of neighboring words, fed to a Maximum Entropy learner. See also CLaC1	63.1	63.1	70.3	70.3
SyntaLex-1	Bagged decision trees using local POS features. See also SyntaLex-3.	62.4	62.4	69.1	69.1
SyntaLex-2	Similar to SyntaLex-1, but using broad context part of speech features.	61.8	61.8	68.4	68.4
Prob2	Similar to Prob1, but invokes only 12 modules.	61.9	61.9	69.3	69.3

(continúa . . .)

Tabla 6.10: Sistemas participantes en la tarea English Lexical Sample de Senseval-3 (continuación)

		Fine		Coarse	
System/Team	Description	P	R	P	R
Duluth-ELSS U.Minnesota (Pedersen)	An ensemble approach, based on three bagged decision trees, using unigrams, bigrams, and co-occurrence features	61.8	61.8	70.1	70.1
UJAEN U.Jaén (García-Vega)	A Neural Network supervised system, using features based on semantic relations from WordNet extracted from the training data	61.3	61.3	69.5	69.5
R2D2 U. Alicante (Vazquez)	A combination of supervised (Maximum Entropy, HMM Models, Vector Quantization, and unsupervised (domains and conceptual density) systems.	63.4	52.1	69.7	57.3
IRST-Ties ITC-IRST (Strapparava)	A generalized pattern abstraction system, based on boosted wrapper induction, using only few syntagmatic features.	70.6	50.5	76.7	54.8
NRC-Coarse	Similar to NRC-Fine; maximizes the coarse score, by training on coarse senses.	48.5	48.5	75.8	75.8
NRC-Coarse2	Similar to NRC-Coarse, with a different threshold for dropping features.	48.4	48.4	75.7	75.7
U.Alicante (Vazquez)	A maximum entropy method and a bootstrapping algorithm ("retraining") with, iterative feeding of training cycles with new high-confidence examples.	78.2	31.0	82.8	32.9

Universitat d'Alacant
Universidad de Alicante

6.2.4 DLSA: English Lexical Sample

El nuevo método de WSD basado en LSA, presentado en el capítulo anterior ha sido evaluado sobre la tarea *"English Lexical Sample"* de Senseval-3. En líneas generales, este método utiliza como base para representar el conocimiento, una matriz [dominios - términos], donde cada columna se corresponde con una categoría semántica de los dominios de WND, y cada fila se corresponde con un término (lema). Para la obtención de la matriz conceptual se utiliza como fuente de información las glosas de WND, ya que, están anotadas con sus correspondientes categorías

semánticas (dominio). Finalmente, tras la obtención de la matriz, el espacio conceptual se reduce a una matriz de 100 dimensiones.

En el proceso de desambiguación se ha utilizado el método DLSA con dos aproximaciones diferentes: una primera aproximación utilizando como matriz conceptual todo el conjunto de palabras con contenido semántico (nombres, verbos, adjetivos y adverbios) y una segunda aproximación restringiendo el tipo de categorías semánticas en la matriz (sólo nombres, sólo verbos o sólo adjetivos). El objetivo de estas dos aproximaciones es determinar si existe alguna interferencia motivada por el uso de contextos más o menos restringidos.

6.2.4.1 Matriz conceptual NVAR.

En esta sección se presentan los resultados obtenidos por el método DLSA utilizando como fuente de información una matriz conceptual construida a partir de todos los nombres, verbos, adjetivos y adverbios de las glosas de WordNet. Para realizar la codificación de la matriz se han obtenido previamente los lemas de las palabras de las glosas de WordNet. Esto se debe a que LSA toma como datos diferentes un nombre en plural y el mismo nombre en singular. Nuestra hipótesis es que ni el tiempo verbal, ni los plurales alteran el contenido semántico de las palabras de un contexto. Una vez obtenida la matriz inicial con 162 dimensiones, cada una de ellas correspondiendo a un dominio de WND, se procede a la reducción de la matriz a únicamente 100 dimensiones. A partir de aquí se han utilizado diferentes heurísticas para determinar el sentido correcto de cada palabra:

- **20 dominios más relevantes:** Sobre los 20 primeros dominios obtenidos por el algoritmo de LSA.
- **10 dominios más relevantes:** Sobre los 10 primeros dominios obtenidos por el algoritmo de LSA.

Pasos para la obtención del sentido correcto DLSA_WSD:

1. **Aplicar LSA sobre el contexto de la palabra ambigua.** Devuelve los dominios con el grado de similitud más elevado respecto al contexto.

2. **Comparar los dominios obtenidos con LSA con los dominios relevantes de cada sentido de la palabra ambigua.** Para cada sentido de la palabra ambigua, se seleccionan aquellos dominios que coinciden con los Dominios Relevantes de ese sentido.
3. **Selección de heurística (para cada posible sentido):**

   3.1. Se suman los valores del Ratio de Asociación (RA) para cada uno de los dominios seleccionados.

   3.2. Se suman los valores de similitud obtenidos con LSA para los dominios seleccionados (esta heurística es la que proporciona mejores resultados).

   3.3 En este tercera heurística, los dominios obtenidos por LSA tienen asociado un valor de similitud, que será más elevado cuanto mayor sea la similitud del contexto con el dominio. Para dar mayor peso a los dominios con mayor valor de similitud se ha optado por realizar el producto de los valores de similitud de LSA por el RA. De esta forma se le da prioridad a los dominios con mayor valor de similitud.

En la Tabla 6.11 se muestran los resultados obtenidos para cada heurística.

		10 dominios		20 dominos	
		Fine	**Coarse**	**Fine**	**Coarse**
Nombres	**Heurística 1**	21.3	33.5	21.5	36.2
	Heurística 2	44.9	52	41.4	48.4
	Heurística 3	21.2	33.4	21.5	36.2
Verbos	**Heurística 1**	35.4	41.9	34.8	42.5
	Heurística 2	51.3	55	49	53.7
	Heurística 3	35.4	41.9	34.8	42.5
Adjetivos	**Heurística 1**	11.1	18.3	9.8	18.3
	Heurística 2	41.8	50.03	37.3	44.4
	Heurística 3	11.1	18.3	9.8	18.3

Tabla 6.11. DLSA aplicado sobre todas las categorías NVAR

Según la Tabla 6.11, la Heurística 2 proporciona los mejores resultados. En este caso, aplicando LSA sobre una matriz concep-

tual con todas las categorías semánticas (nombres, verbos, adjetivos y adverbios), se alcanza una precisión de alrededor de 45 % en el caso de evaluar el sistema con granularidad fina y de un 55 % en el caso de evaluar el sistema con granularidad gruesa. Los valores obtenidos también demuestran que la utilización de los 10 primeros dominios como fuente de información semántica es suficiente para alcanzar buenos resultados. En cambio, la utilización de los 20 primeros dominios empeora los resultados gradualmente.

6.2.4.2 Matriz conceptual N-V-A.

Los resultados obtenidos tras la evaluación de los nombres (aproximación matriz con sólo nombres) se reflejan en la Tabla 6.12, donde se muestra el grado de precisión/recall obtenidos para cada palabra en concreto. En todos nuestros experimentos, siempre se han contestado todas las instancias, por tanto, precisión y recall alcanzan el mismo resultado.

Para los verbos y los adjetivos, los resultados individuales obtenidos (aproximación matriz sólo verbos y sólo adjetivos) son los mostrados en las Tablas 6.13 y 6.14.

En la Tabla 6.15 se muestran los resultados obtenidos para cada heurística.

Al igual que sucedía en el experimento anterior, los mejores resultados según la Tabla 6.15 son los de la Heurística 2. En este caso se han realizado tres matrices conceptuales distintas, una por cada categoría (N, V, A o R). En este caso, los resultados obtenidos en este último experimento, mejoran los anteriores. Es decir, especializando las matrices conceptuales y restringiendo el uso de categorías léxicas, se pueden mejorar los resultados en el proceso de desambiguación. De esta forma, si se trata de desambiguar un nombre, la información contextual perteneciente a los nombres que lo rodean proporciona un mejor indicativo de sus relaciones semánticas, que todas las demás palabras que lo rodean.

Palabra	Precisión/Recall
argument	53.27
arm	67.74
atmosphere	57.79
audience	51.51
bank	32.5
degree	61.41
difference	40.35
difficulty	18.18
disc	27
image	37.68
interest	23.65
judgment	15.62
organization	39.28
paper	4.28
party	18.91
performance	26.74
plan	82.14
shelter	32.98
sort	82.89
source	68.96

Tabla 6.12. Resultados ELS sobre nombres

6.2.4.3 Comparativa con otros sistemas.

En la tarea *"English Lexical Sample"*, participaron tanto sistemas supervisados como sistemas no supervisados. Los primeros obtuvieron mejores resultados alcanzando el mejor sistema un 72.9 % de precisión y cobertura (Grozea (2004)). Con respecto a los sistemas no supervisados, el mejor sistema obtuvo un 66.1 % de precisión (Ramakrishnan et al. (2004)). Dado que DLSA es un sistema no supervisado, realizaremos la comparativa de los resultados obtenidos con estos últimos. En la Tabla 6.16 se muestran los resultados obtenidos para cada uno de los sistemas participantes junto con una breve descripción de cada uno de ellos.

La heurística que mejores resultados proporciona para nuestro método DLSA es la heurística 2, la cual, selecciona el sentido más adecuado a partir de la intersección de los dominios obtenidos por

Palabra	Precisión/Recall
activate	8.84
add	48.85
appear	45.03
ask	29.35
begin	60.52
climb	78.12
decide	51.61
eat	97.70
encounter	53.12
expect	74.35
express	88.88
hear	40.62
lose	81.81
mean	30
miss	21.53
note	95.52
operate	16.66
play	17.64
produce	89.24
provide	72.46
receive	44.44
remain	37.14
rule	50
smell	41.17
suspend	13.17
talk	89.39
treat	47.36
use	100
wash	16.45
watch	98.03
win	50
write	19.09

Tabla 6.13. Resultados ELS sobre verbos

Universitat d'Alacant
Universidad de Alicante

Palabra	Precisión/Recall
different	53.19
hot	80.95
important	21.05
simple	27.77
solid	17.03

Tabla 6.14. Resultados ELS sobre adjetivos

		10 dominios		20 dominios	
		Fine	**Coarse**	**Fine**	**Coarse**
Nombres	**Heurística 1**	23.8	38	26.2	42
	Heurística 2	45.8	53.3	44	50.4
	Heurística 3	23.8	38.1	26.1	41.8
Verbos	**Heurística 1**	36.2	41.8	37.6	43.1
	Heurística 2	53.2	58	52.7	57.6
	Heurística 3	36.4	41.8	38	43.4
Adjetivos	**Heurística 1**	11.1	19	10.5	17.6
	Heurística 2	45.1	54.9	41.8	49.7
	Heurística 3	10.5	18.3	11.1	19

Tabla 6.15. DLSA aplicado sobre cada categoría por separado

LSA y los dominios relevantes del vector de sentidos, utilizando únicamente los valores de similitud obtenidos por LSA.

El resultado global tras la evaluación sobre el corpus de test de Senseval-3, posiciona nuestro sistema en cuarto lugar con respecto al resto de sistemas no supervisados participantes en la tarea English Lexical Sample.

6.2.5 SenseDiscrim: Spanish Lexical Sample

El sistema SenseDiscrim basado en reglas lingüísticas aplicadas sobre corpus y en la obtención de conjuntos de discriminadores de sentidos sobre WordNet, se ha evaluado sobre el corpus de test de Senseval-3 en la tarea Spanish Lexical Sample.

Dado que este sistema actualmente se ha desarrollado para la desambiguación de nombres, es necesaria la combinación con otro

		Fine		Coarse	
Sistema	Descripción	P	R	P	R
wsdiit IIT Bombay (Ramakrishnan et al.)	Utiliza la medida de similitud de Lesk entre los contextos de palabras ambiguas y definiciones de diccionarios.	66.1	65.7	73.9	74.1
Cymfony (Niu)	Utiliza un modelo de máxima entropía para clustering no supervisado, utilizando palabras vecinas y estructuras sintácticas como características.	56.3	56.3	66.4	66.4
Prob0 Cambridge U. (Preiss)	Combina dos módulos no supervisados, utilizando información básica de POS-tagging e información de frecuencia.	54.7	54.7	63.6	63.6
DLSA H2 University of Alicante (svazquez)	Utiliza LSA combinada con Dominios Relevantes. Heurística 2.	48.9	48.9	54.2	54.2
clr04-ls CL Research (Litkowski)	Utiliza una serie de propiedades (sintácticas, semánticas, patrones de subcategorización, otra información léxica).	45.0	45.0	55.5	55.5
CIAOSENSO U. Genova (Buscaldi)	Combina la densidad conceptual con la frecuencia de palabras y la información proporcionada por dominios.	50.1	41.7	59.1	49.3
KUNLP Korea U. (Seo)	Selecciona el sentido de las palabras utilizando sustitutos a través de la jerarquía de WordNet (antónimos, hiperónimos, etc). La selección se hace a partir de la co-ocurrencia de palabras en un corpus.	40.4	40.4	52.8	52.8
Duluth-SenseRelate U.Minnesota (Pedersen)	Asigna el sentido mejor relacionado con los posibles sentidos de las palabras vecinas. Se utilizan las glosas de WordNet para medir la similitud entre sentidos.	40.3	38.5	51.0	48.7
DFA-LS-Unsup UNED (Fernandez)	Combina tres heurísticas: similitud entre sinónimos y el contexto, de acuerdo a la medida de la información mutua; patrones léxico-sintácticos a partir de las glosas de WordNet y la heurística del primer sentido.	23.4	23.4	27.4	27.4

Tabla 6.16. Sistemas no supervisados en la tarea ELS de SENSEVAL-3

Universitat d'Alacant
Universidad de Alicante

sistema para responder a las instancias referentes a verbos y adjetivos. En nuestro caso, se ha optado por la utilización del sistema supervisado de Máxima Entropía (Suárez (2004)) que ha demostrado obtener una alta precisión en el proceso de desambiguación.

En la tarea Spanish Lexical Sample se debían desambiguar 21 nombres: arte, autoridad, banda, canal, circuito, columna, corazón, corona, gracia, grano, hermano, letra, masa, mina, naturaleza, operación, órgano, partido, pasaje, programa y tabla. De estos 21 nombres sólo se realizó un análisis parcial de los 13 nombres mostrados en la Tabla 6.17.

El objetivo principal de la evaluación del sistema SenseDiscrim, es demostrar que mediante la elección de un determinado número de patrones de los que se puede extraer información paradigmática y de un conjunto de discriminadores de sentidos, se puede obtener una alta precisión que supera a la mayoría de sistemas actuales.

Para la adquisición de información paradigmática a partir de patrones sintagmáticos se ha utilizado el corpus EFE sobre noticias en español, junto con un umbral= 5 de frecuencia mínima para la extracción de información.

El procedimiento seguido para la desambiguación de nombres es el siguiente:

1. Se identifican los patrones sintácticos con un filtro de frecuencia mínimo de 5.
2. Se adquiere la información paradigmática de los patrones identificados utilizando el corpus EFE como corpus de búsqueda.
3. Se utiliza el algoritmo de Prueba de Conmutabilidad estableciendo un sentido para cada patrón identificado en el paso previo.
4. Para cada patrón: se interseccionan las propuestas de sentidos a partir de la información sintagmática y de la información paradigmática. El sentido propuesto por la mayoría de los patrones es seleccionado. En caso, de no existir acuerdo, se da preferencia al sentido más frecuente en WordNet.

El análisis posterior de los resultados demuestra que para la mitad de los nombres de la tarea de *"Spanish Lexical Sample"*, no existe información suficiente en el corpus EFE que permita

Universitat d'Alacant
Universidad de Alicante

Palabra	Ocurrencias	Contestadas	Correctas	Cobertura	Precisión
autoridad	132	38	35	28.79 %	92.11 %
canal	131	21	21	16.03 %	100 %
circuito	132	3	1	1.52 %	50 %
corona	64	0	0	0 %	0 %
gracia	38	0	0	0 %	0 %
grano	61	2	0	3.28 %	0 %
hermano	66	0	0	0 %	0 %
masa	85	0	0	0 %	0 %
naturaleza	128	0	0	0 %	0 %
partido	66	17	14	25.76 %	82.35 %
pasaje	111	0	0	0 %	0 %
programa	133	26	23	19.55 %	88.46 %
tabla	64	0	0	0 %	0 %

Tabla 6.17. Resultados del sistema SenseDiscrim para los nombres de la tarea Spanish Lexical Sample de SENSEVAL-3

identificar patrones con un grado de fiabilidad elevado. En cambio, para el resto de nombres identificados, la precisión obtenida es muy elevada, tal y como se pretendía demostrar. En nuestro caso, mediante la utilización de patrones y la combinación de la información paradigmática y el conjunto de discriminadores de sentidos de WordNet, es posible realizar una desambiguación con una alta eficacia.

Universitat d'Alacant
Universidad de Alicante

6.2.5.1 Evaluación de los resultados.

Los resultados obtenidos tras la evaluación fueron de un 84 % de precisión y un 47 % de cobertura. Cabe destacar que los resultados en cuanto a precisión son excelentes en detrimento de una baja cobertura. Esto es debido en gran parte, a la escasez de corpus en español, lo que supone un impedimento para la extracción de información paradigmática.

La Tabla 6.18 muestra los resultados obtenidos por los sistemas que participaron en la tarea *"Spanish Lexical Sample"* de SENSEVAL-3.

Sistema	Prec.	Recall	Cover.	$F_{\beta=1}$
IRST	84.20 %	84.20 %	100.0 %	84.20
UA-SRT	84.00 %	84.00 %	100.0 %	84.00
UMD	82.48 %	82.48 %	100.0 %	82.48
UNED	81.76 %	81.76 %	100.0 %	81.76
SWAT	79.45 %	79.45 %	100.0 %	79.45
D-SLSS	74.29 %	75.02 %	100.0 %	74.65
CSUSMCS	67.84 %	67.82 %	99.9 %	67.83
UA-NSM	61.93 %	61.93 %	100.0 %	61.93
UA-NP	84.31 %	47.27 %	56.1 %	60.58
MFC	67.72 %	67.72 %	100.0 %	67.72
COMB	85.98 %	85.98 %	100.0 %	85.98

Tabla 6.18. Resultados de los sistemas participantes en la tarea Spanish Lexical Sample Senseval-3

Los resultados mostrados en la Tabla 6.18 se encuentran ordenados según la medida $F_{\beta=1}$. Nuestro sistema alcanza la mayor precisión, sin embargo, obtiene una cobertura muy baja debido a la escasez de corpus de los que extraer patrones e información paradigmática.

El mejor sistema en esta tarea fue IRST (Strapparava et al. (2004)) que utilizaba SVM como algoritmo de aprendizaje. Este sistema obtuvo los mejores resultados para las palabras con menos ejemplos por sentido.

Los dos últimos sistemas mostrados en la tabla corresponden a un sistema utilizado como baseline MFC (Most Frequent Sense Clasifier) y un sistema de votación COMB que combina las respuestas de los mejores sistemas en la tarea.

6.2.6 Web People Search

En la cuarta edición de Senseval se presentó una nueva tarea denominada "Web People Search" (WePS), esta tarea tiene como objetivo la detección y clasificación de distintos documentos (páginas web) a partir de los nombres propios que aparecen en ellos. La dificultad de esta tarea viene determinada por la existencia de nombres propios ambiguos que aparecen en diferentes

contextos. Es por tanto necesario distinguir entre los distintos contextos de los documentos y establecer para cada contexto los nombres propios relacionados con ellos.

Debido a la necesidad de utilizar la información contextual como base para la detección y distinción de diferentes entidades, se ha considerado viable la utilización de la técnica de LSA para esta tarea. En concreto, se ha utilizado LSA para la agrupación de contextos similares a partir de la información semántica contenida. Tras la obtención de los contextos similares se ha utilizado una técnica de clustering para la creación de conjuntos disjuntos de documentos.

El sistema presentado en esta tarea consta de varios módulos. A continuación se describen cada uno de ellos:

- **Módulo de preproceso.** El primer módulo se encarga de realizar el preproceso de los documentos de entrada. Dado que todos los documentos son páginas web, existen etiquetas específicas del lenguaje HTML y código Javascript que no deben tenerse en cuenta. Por tanto, mediante un proceso de detección y eliminación de elementos propios del lenguaje HTML se obtiene únicamente el texto comprendido entre las etiquetas <title> </title> y las etiquetas <body> </body>.
- **Módulo de extracción de información contextual.** En este módulo se utilizan los resultados obtenidos en el módulo de preproceso para la extracción de información relevante que ayude a identificar correctamente los diferentes contextos. Este módulo se divide en cuatro sub-procesos:
  - **Detección de nombres.** Todos los nombres propios del contexto (personas, organizaciones, lugares, etc), son detectados y extraídos. Para esta tarea se ha utilizado la arquitectura GATE[6] (Cunningham (2002), Cunningham (2005)) que integra un módulo de detección de entidades. El objetivo de este sub-módulo es obtener los nombres propios de las diferentes categorías y detectar sus ocurrencias en el resto de documentos. De esta forma, documentos que compartan las mismas entidades pueden hacer referencia al mismo in-

[6] http://gate.ac.uk/

dividuo. Este sub-módulo retorna como salida una matriz de valores binarios, donde 1 significa que los documentos comparados comparten más de la mitad de sus entidades y 0 cualquier otro caso.

- **Identificación de enlaces web.** Para cada documento se han extraído los enlaces comprendidos entre las etiquetas <a href> . Dado que los enlaces detectados son muy específicos se ha empleado una función que extrae la raíz general de cada uno de los enlaces. Por ejemplo, la dirección d_1 `http://www.cs.ualberta.ca/~lindek/index.htm` se transformará en `http://www.cs.ualberta.ca/~lindek`. De esta forma si dos direcciones cualquiera d_1 y d_2, comparten la misma raíz $d_1 \bigcap d_2$ se consideran la misma dirección. El nivel de profundidad seleccionado de cada enlace es 3 niveles como máximo y 2 niveles como mínimo. La salida de este sub-módulo es una matriz de valores binarios con 1 si el par de documentos comparten más de 3 enlaces y 0 en otro caso.
- **Detección de títulos.** Para cada documento se han extraído los títulos comprendidos entre las etiquetas <title> </title>. Éstos se han introducido en una matriz de unigramas utilizada como entrada para un sistema automático de clustering SenseClusters[7]. Mediante un criterio de clustering con parada automática se han agrupado los diferentes documentos de acuerdo al contexto de los títulos. Del resultado obtenido se ha generado una nueva matriz de valores binarios con 1 para los pares de documentos situados en el mismo cluster y 0 en otro caso.
- **Tratamiento del cuerpo de la página web.** La parte de texto comprendida entre las etiquetas <body> </body> ha sido tratada para extraer las categorías sintácticas de las palabras usando Tree Tagger[8]. El resultado de este sub-módulo es la anotación del texto con la información sintáctica de cada palabra: "water#v the#det flowers#n and#conj pass#v me#pron the#det glass#n of#prep water#n". El objetivo

[7] `http://www.d.umn.edu/~tpederse/senseclusters.html`
[8] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

de esta transformación es servir de entrada al módulo de LSA para poner tener en cuenta las categorías sintácticas de las palabras y poder construir así una matriz conceptual más precisa, ya que, no es lo mismo encontrar una palabra actuando como nombre ("water#n"), que actuando como verbo ("water#v").

- **Módulo de clustering.** El último módulo es el módulo de clasificación de documentos en sus correspondientes clusters. En este módulo se realizan tres sub-tareas.
  - **LSA.** Utilizando LSA, a partir de la codificación obtenida en el sub-módulo de tratamiento del cuerpo de los documentos, se construye la matriz conceptual. En esta matriz las filas representan palabras de la colección de documentos y las columnas representan los documentos (páginas web) y las celdas contienen le frecuencia de ocurrencia de cada palabra en cada documento. A continuación se ha reducido la matriz a 300 dimensiones para evitar el ruido causado por información irrelevante. Finalmente, la salida de este sub-módulo es una matriz que representa el grado de similitud entre los diferentes documentos.
  - **Combinación de información contextual.** En este sub-módulo se han combinado los resultados de la extracción de información referente a entidades, títulos, enlaces y cuerpo de los documentos. Esta información se ha introducido en una nueva matriz de 100x400 dimensiones. Las filas se corresponden con el número de documentos y las columnas representan los valores obtenidos para las entidades, títulos, enlaces y cuerpo de los documentos. Esta matriz es la entrada a un algoritmo de clustering denominado K-means[9] que determina el clustering de documentos a partir de esa información.
  - **K-means.** Para realizar el clustering de N páginas web en K conjuntos disjuntos S_j que contienen N_j puntos de datos, se utiliza la minimización del cuadrado de sumas

[9] `http://www.cs.waikato.ac.nz/ml/weka/`

$J = \sum_{j=1}^{K} \sum_{n \in S_j} |x_n - mu_j|^2$, donde x_n es un vector que representa el n_{avo} punto de dato y mu_j es el centroide geométrico de los puntos de datos en S_j. La matriz a partir de la cual se realiza el clustering incluye la información del título, enlaces, entidades y cuerpo de las páginas web. En la implementación de K-means no existe un criterio de parada automático (Witten y Frank (1999)) por lo que se estableció manualmente.

La arquitectura del sistema WePS se muestra en la Figura 6.2.

6.2.6.1 Evaluación de los sistemas de la tarea WePS.

Los datos utilizados para la evaluación de los sistemas en la tarea WePS fueron extraídos de Wikipedia, de las personas que participaron en ACL06 y del corpus Web03 (Mann (2006)) que contiene 32 nombres aleatorios extraídos del US Census. En la Tabla 6.19 se muestra la relación de los diferentes nombres propios con su nivel de ambigüedad.

Las medidas utilizadas para la evaluación de los distintos sistemas fueron "Purity" e "Inverse Purity". La medida de "Purity" está relacionada con la medida de Precisión, muy utilizada en Recuperación de Información. Esta medida se centra en la frecuencia de las categorías más comunes en cada cluster y da mayor puntuación a los sistemas cuya clasificación introduce menos ruido en los clusters. Siendo C el conjunto de clusters a ser evaluados, L el conjunto de categorías (anotadas manualmente) y n el número de elementos clasificados, la medida "Purity" se obtiene según la Fórmula 6.3.

$$Purity = \sum_{i} \frac{|C_i|}{n} \max Precision(C_i, Lj) \tag{6.3}$$

Universidad de Alicante

Nombre	Entidades	Documentos	Descartados
Wikipedia names			
Arthur Morgan	19	100	52
James Morehead	48	100	11
James Davidson	59	98	16
Patrick Killen	25	96	4
William Dickson	91	100	8
George Foster	42	99	11
James Hamilton	81	100	15
John Nelson	55	100	25
Thomas Fraser	73	100	13
Thomas Kirk	72	100	20
Average	56.50	99.30	17,50
ACL06 Names			
Dekang Lin	1	99	0
Chris Brockett	19	98	5
James Curran	63	99	9
Mark Johnson	70	99	7
Jerry Hobbs	15	99	7
Frank Keller	28	100	20
Leon Barrett	33	98	9
Robert Moore	38	98	28
Sharon Goldwater	2	97	4
Stephen Clark	41	97	39
Average	31.00	98.40	12,80
US Census Names			
Alvin Cooper	43	99	9
Harry Hughes	39	98	9
Jonathan Brooks	83	97	8
Jude Brown	32	100	39
Karen Peterson	64	100	16
Marcy Jackson	51	100	5
Martha Edwards	82	100	9
Neil Clark	21	99	7
Stephan Johnson	36	100	20
Violet Howard	52	98	27
Average	50.30	99.10	14.90
Global average	45.93	98.93	15.07

Tabla 6.19. Nombres ambiguos en WePS

Universitat d'Alacant
Universidad de Alicante

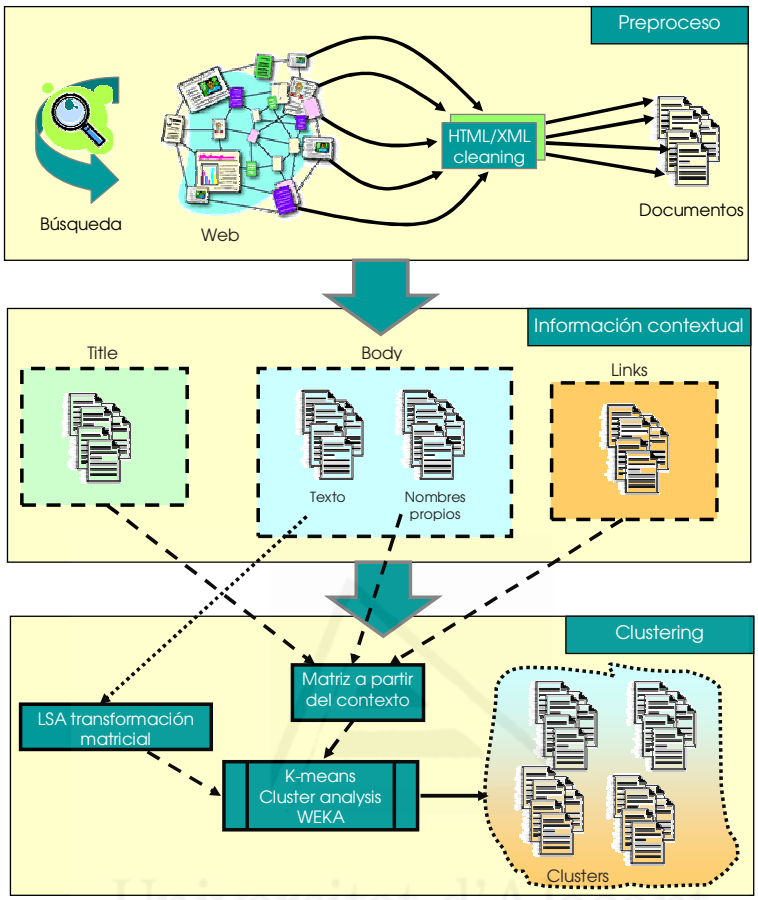


Figura 6.2. Arquitectura sistema WePS

Universitat d'Alacant
Universidad de Alicante

Donde la precisión de un cluster C_i para una categoría determinada L_j se define como:

$$Precision = (C_i, L_j) = \frac{|C_i \bigcap L_j|}{|C_i|} \tag{6.4}$$

La medida "Inverse Purity" se centra en el cluster con máximo "recall" para cada categoría, dando mayor puntuación a los resultados que proporcionan mayor número de elementos para ca-

da categoría en su correspondiente cluster. Esta medida se define según la Fórmula 6.5:

$$InversePurity = \sum_i \frac{|L_i|}{n} \max Precision(L_i, C_j) \tag{6.5}$$

Para la clasificación final de los sistemas, se utilizó la medida armónica $F_{\alpha=0,5}$. Esta medida se define según la Fórmula 6.6:

$$F = \frac{1}{\alpha \frac{1}{Purity} + (1-\alpha)\frac{1}{InversePurity}} \tag{6.6}$$

Además se añadió otro valor a α para dar mayor importancia a la "Inverse Purity", $\alpha = 0{,}2$. La idea es que para un buscador web, debería ser más fácil desechar unas pocas páginas web incorrectas en un cluster que contenga toda la información necesaria, que tener que obtener la información a partir de diversos clusters. Por lo tanto, el alcanzar un valor elevado en la "Inverse Purity" debería tenerse también en cuenta a la hora de evaluar los diferentes sistemas.

La Tabla 6.20 muestra los resultados obtenidos tras la evaluación de nuestro sistema.

Como se observa en la Tabla 6.20, la media de efectividad de nuestro sistema está alrededor del 56 %. Con respecto a los otros participantes de la tarea WePS nuestro sistema se sitúa en la décima posición de entre dieciséis participantes (Ver Tabla 6.21).

Tras el análisis de los resultados obtenidos se detectaron algunas limitaciones a la hora de asignar correctamente los clusters. Por ejemplo, existían muchas páginas web que no contenían apenas información entre las etiquetas <body> </body>, lo que dificultaba el correcto funcionamiento de LSA, ya que, el resultado de calcular la similitud de un documento sin información en el cuerpo de la página web respecto a otros, daba como resultado 0.

Otra limitación viene dada por la diferencia de tamaño de los contextos de las distintas páginas web. Una consecuencia de esta variedad es que LSA obtiene peores resultados si los contextos no

Name	Purity	Inverse Purity	F α=0.5	F α=0.2
Mark Johnson	0,55	0,74	0,63	0,69
Sharon Goldwater	0,96	0,23	0,37	0,27
Robert Moore	0,36	0,67	0,47	0,57
Leon Barrett	0,62	0,51	0,56	0,52
Dekang Lin	0,99	0,43	0,60	0,49
Stephen Clark	0,52	0,75	0,62	0,69
Frank Keller	0,38	0,67	0,48	0,58
Jerry Hobbs	0,54	0,63	0,58	0,61
James Curran	0,53	0,61	0,57	0,59
Chris Brockett	0,73	0,40	0,51	0,44
Thomas Fraser	0,66	0,57	0,61	0,58
John Nelson	0,68	0,76	0,72	0,74
James Hamilton	0,56	0,60	0,58	0,59
William Dickson	0,59	0,78	0,67	0,73
James Morehead	0,36	0,64	0,46	0,56
Patrick Killen	0,56	0,69	0,62	0,66
George Foster	0,46	0,70	0,56	0,64
James Davidson	0,58	0,71	0,64	0,68
Arthur Morgan	0,77	0,47	0,59	0,51
Thomas Kirk	0,26	0,90	0,41	0,60
Patrick Killen	0,56	0,69	0,62	0,66
Harry Hughes	0,66	0,54	0,59	0,56
Jude Brown	0,64	0,63	0,64	0,63
Stephan Johnson	0,56	0,80	0,66	0,73
Marcy Jackson	0,40	0,73	0,52	0,63
Karen Peterson	0,56	0,72	0,63	0,68
Neil Clark	0,68	0,36	0,47	0,40
Jonathan Brooks	0,53	0,76	0,63	0,70
Violet Howard	0,58	0,75	0,65	0,71
Global average	0,58	0,64	0,58	0,60

Tabla 6.20. Resultados evaluación WePS

Universitat d'Alacant
Universidad de Alicante

son de un tamaño semejante. En un futuro se pretende trabajar con una ventana de tamaño específico para cada contexto.

Además, en WePS el número de individuos que comparten el mismo nombre es desconocido. Por tanto, el establecimiento del número de clusters es muy complicado y en el caso de K-means que no dispone de un criterio de parada, se debe hacer de forma

manual. En nuestro experimento, el establecimiento del número de clusters se hizo evaluando los resultados obtenidos para diferentes rangos sobre el corpus de entrenamiento. Los resultados demostraron que el número de clusters idóneo para un correcto funcionamiento se establecía entre 25 y 50.

En la Tabla 6.21 se pueden consultar los resultados obtenidos por el resto de participantes en la tarea.

		Macro-averaged Scores			
		F-measures			
rank	team-id	$\alpha = 0{,}5$	$\alpha = 0{,}2$	Pur	Inv_Pur
1	CU COMSEM	0.78	0.83	0.72	0.88
2	IRST-BP	0.75	0.77	0.75	0.80
3	PSNUS	0.75	0.78	0.73	0.82
4	UVA	0.67	0.62	0.81	0.60
5	SHEF	0.66	0.73	0.60	0.82
6	FICO	0.64	0.76	0.53	0.90
7	UNN	0.62	0.67	0.60	0.73
8	ONE-IN-ONE	0.61	0.52	1.00	0.47
9	AUG	0.60	0.73	0.50	0.88
10	SWAT-IV	0.58	0.64	0.55	0.71
11	UA-ZSA	0.58	0.60	0.58	0.64
12	TITPI	0.57	0.71	0.45	0.89
13	JHU1-13	0.53	0.65	0.45	0.82
14	DFKI2	0.50	0.63	0.39	0.83
15	WIT	0.49	0.66	0.36	0.93
16	UC3M 13	0.48	0.66	0.35	0.95
17	UBC-AS	0.40	0.55	0.30	0.91
18	ALL-IN-ONE	0.40	0.58	0.29	1.00

Tabla 6.21. Resultados evaluación sistemas WePS

Los resultados presentados en la Tabla 6.21 responden a la medida del macro-promedio[10] en lugar de a la medida del micro-promedio[11]. Se optó por estas medidas porque el macro-promedio

[10] El macro-promedio viene determinado por calcular el valor de F para cada persona y después calcular la media entre los resultados obtenidos

[11] El micro-promedio viene determinado por las medidas de Purity e Inverse Purity sobre todas las instancias, para luego calcular el valor de F sobre esos resultados

tiene una interpretación más clara: si la medida de evaluación es F, entonces se debería calcular F para cada nombre de persona y entonces calcular la media de todos los valores de F obtenidos.

6.3 Participación en iCLEF

Dentro del marco de la competición CLEF (Cross-Language Evaluation Forum) para la evaluación de sistemas en espacios multilingües, se participó en la tarea específica iCLEF (Gonzalo y Oard (2004)). En esta tarea (cross-language search system), el objetivo era determinar la forma de proporcionar la mejor asistencia a distintos usuarios que formulaban preguntas en su lengua materna y obtenían respuestas en otra lengua distinta.

El conjunto de preguntas proporcionadas para esta tarea fueron extraídas de la tarea de Question Answering del CLEF 2004, con el fin de comparar los resultados obtenidos por sistemas automáticos y los obtenidos con los experimentos interactivos.

En nuestro caso (Navarro et al. (2004)), el sistema proporcionaba varias respuestas (50 párrafos) para una misma pregunta, junto con una serie de indicadores que establecían grados de similitud entre cada par (pregunta-respuesta). Las preguntas estaban formuladas en español y las respuestas en inglés.

Para lograr el objetivo de asistir a los usuarios a localizar la respuesta correcta, nos centramos en dos ideas:

1. **El tipo de información mostrada al usuario.** Debe ser suficiente para la localización de la respuesta correcta, dado que el usuario no conoce de antemano la respuesta a cada pregunta. Es el usuario el que debe decidir si la respuesta se encuentra en el párrafo mostrado o no. Por tanto, no sólo se debe mostrar la respuesta de forma explícita, sino el suficiente contexto para poder extraer la respuesta correcta.
2. **Cómo se muestra la información al usuario.** Concretamente en qué lengua se le muestra la información al usuario. Si los usuarios no conocen la lengua de los párrafos que muestran las respuestas, se les debe proporcionar alguna información extra para localizar la respuesta correcta.

Universidad de Alicante

A partir de estas premisas, usuarios con poco conocimiento del inglés podrían establecer la respuesta correcta a partir de resultados obtenidos en inglés y no en su lengua materna. Además, mediante el uso de este tipo de sistemas, se podría evitar la traducción de grandes volúmenes de datos a diferentes idiomas para satisfacer la demanda de información a partir de otras lenguas.

6.3.1 Desarrollo de los experimentos

Con el objetivo de asistir a los usuarios en su búsqueda de información, se han seguido tres pasos para realizar los experimentos.

1. **Formulación de la pregunta y traducción automática.** Se extrajeron preguntas en español de la colección del CLEF 2004 y fueron traducidas con un sistema automático de traducción a Inglés [12].
2. **Extracción de pasajes relevantes.** Para localizar los pasajes relevantes de la colección de documentos en inglés se utilizó un sistema de recuperación de información automático (Llopis (2003)). Este sistema extrae pasajes relacionados con la pregunta y los ordena de mayor relevancia a menor relevancia. Concretamente, el tamaño de cada pasaje extraído fue de 5 frases, tamaño más que suficiente para localizar una posible respuesta.
3. **Interacción con el usuario y localización de la respuesta.** Las preguntas en español y los pasajes en inglés se mostraron a los usuarios a través de una página web. Los usuarios debían de localizar la respuesta correcta entre los diferentes pasajes mostrados. Entonces debían seleccionar la respuesta (cadena de caracteres) y el pasaje donde aparecía.

El problema de esta tarea, tal y como se ha mencionado anteriormente, es la falta de conocimiento por parte del usuario de la lengua en la que se muestran los pasajes. Es por tanto necesaria la incorporación de cierta información que ayude al usuario a determinar el pasaje y la respuesta correcta para cada pregunta.

[12] `http://babelfish.yahoo.com/`

Universitat d'Alacant Universidad de Alicante

Para ello, se desarrollaron dos métodos: uno basado en etiquetas semánticas y otro basado en patrones sintáctico-semánticos (SSP).

6.3.1.1 Método interactivo I: Dominios Relevantes.

Este primer método utiliza el recurso léxico Dominios Relevantes para ayudar a la localización de las respuestas por parte de los usuarios. Como ya se explicó en el capítulo anterior los Dominios Relevantes (extraídos a partir de WordNet Domains), son aquellos dominios o etiquetas semánticas más representativas de un palabra. En este caso, nuestra hipótesis es que si sabemos los dominios relevantes de la pregunta y los dominios relevantes de las respuestas, podemos reducir la colección de pasajes a aquellos que compartan ciertos dominios. De esta forma, la respuesta correcta será localizada con mayor facilidad y con una alta probabilidad en aquellos pasajes que compartan la mayor cantidad de dominios relevantes respecto a la pregunta en cuestión.

Un ejemplo de la información proporcionada a los usuarios, se muestra en la Figura 6.3. En esta captura de la página web mostrada a cada usuario se aprecian: la pregunta en cuestión, el pasaje seleccionado, los dominios relevantes de la pregunta y los dominios relevantes del pasaje.

Para la pregunta "¿Quién es el director gerente de FIAT?", los dominios asociados son: Administration, Economy y Transport. Y de entre los dominios mostrados para el pasaje encontramos Economy y Transport entre los cinco primeros. Por tanto, este pasaje podría ser candidato de contener la respuesta correcta.

Además de proporcionar información útil acerca de los dominios involucrados en cada consulta, el orden de los pasajes obtenido por el sistema de recuperación fue alterado. De forma que aquellos pasajes con mayor similitud respecto a la pregunta, fueron mostrados previamente a aquellos cuya similitud era menor.

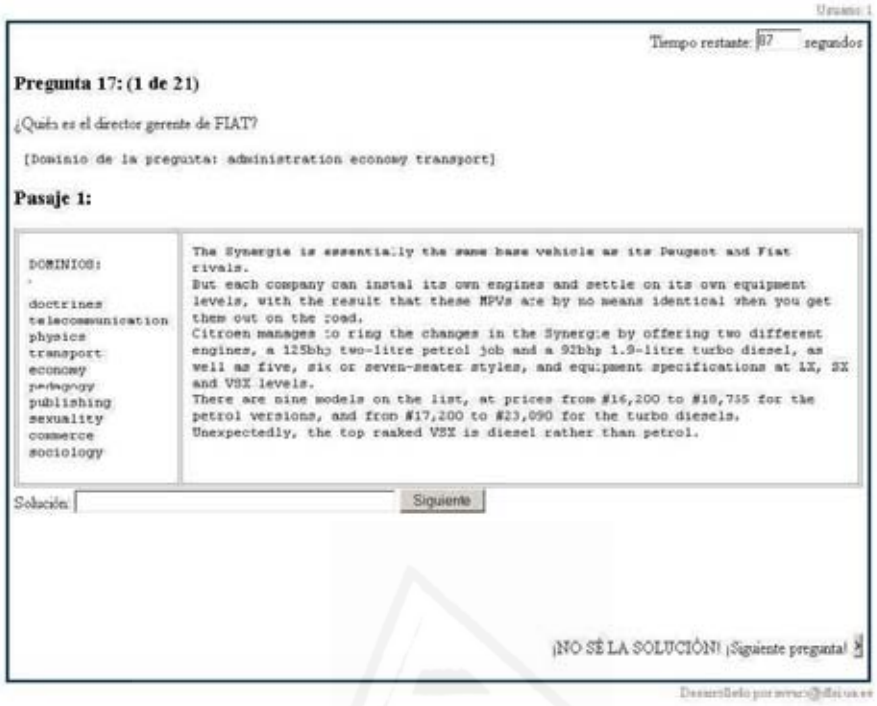


Figura 6.3. Página web interactiva para dominios relevantes

6.3.1.2 Método interactivo II: Patrones sintáctico-semánticos.

El segundo método está basado en patrones sintáctico-semánticos. Con este método se muestra al usuario una serie de patrones (SSP) junto con los pasajes en inglés, donde cada patrón está formado por los verbos y los nombres principales. La hipótesis en este caso es determinar si esta información permite al usuario decidir si el pasaje mostrado contiene la respuesta a la pregunta formulada. Intuitivamente cuando un usuario busca la respuesta a una determinada pregunta en una porción de texto, éste presta más atención a los nombres y verbos, intentando localizar verbos o nombres similares a los de la pregunta. Con los patrones, los verbos y nombres principales son extraídos automáticamente, facilitando esta tarea a los usuarios.

Desde el punto de vista teórico un patrón sintáctico-semántico está formado por tres componentes:

1. Un verbo con su sentido o sentidos.
2. El marco de subcategorización de ese sentido.
3. Las preferencias de selección de cada argumento.

Dado que obtener de forma automática esta información es un proceso muy costoso, el sistema utiliza una versión de SSP's menos compleja. En este nuevo modelo, el verbo se representa por la palabra y su sentido o sentidos, el marco de subcategorización se representa por el nombre principal de cada argumento (si el argumento es una cláusula, en lugar del nombre se utilizará el verbo) y las preferencias de selección de cada argumento se representarán por el sentido o los sentidos de los nombres principales.

En la Figura 6.4 se muestra una captura de la página web mostrada al usuario. En esta captura se aprecia por una parte la pregunta, el pasaje y los SSP's asociados al pasaje.

Figura 6.4. Página web interactiva para patrones SSP

Utilizando los SSP's, sólo la información más importante de cada frase es mostrada al usuario: los verbos y los nombres principales y las relaciones sintáctico-semánticas existentes entre ellos.

6.3.2 Resultados

En la Figura 6.5 se muestran los resultados obtenidos por los usuarios con cada método interactivo. Como se puede observar los usuarios obtuvieron resultados similares con ambos métodos. Únicamente se aprecia una pequeña mejora de 0.015 obtenida por el método de patrones.

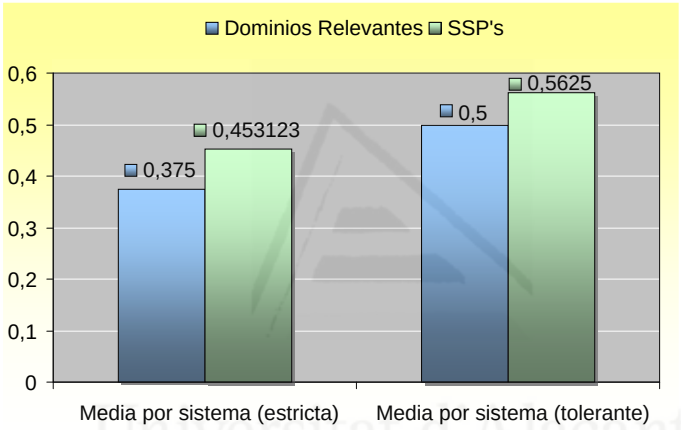


Figura 6.5. Media genérica

Universitat d'Alacant
Universidad de Alicante

Las medidas de efectividad de los sistemas son las mismas utilizadas en la tarea de CL-QA (Cross Language Question Answering). La media estricta son las respuestas correctas obtenidas a partir de pasajes que contienen la respuesta. Y la media tolerante, son las respuestas correctas obtenidas por los usuarios independientemente de encontrarse en el pasaje correcto.

6.3.2.1 Media por usuario.

Las Figuras 6.6 y 6.7 representan la media obtenida por cada usuario. La Figura 6.6 muestra las respuestas correctas encontradas por el usuario en un pasaje que realmente contiene la respuesta (estricta).

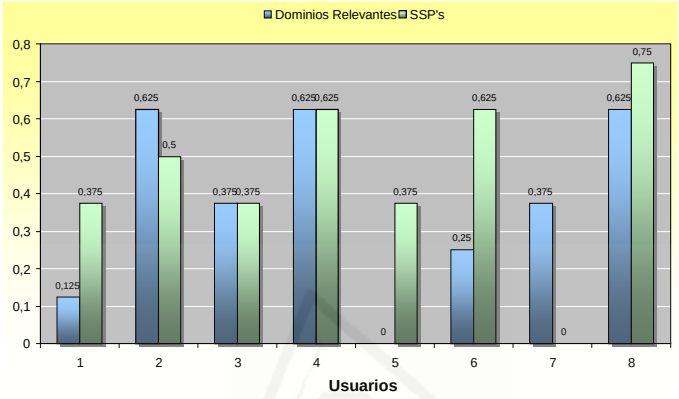


Figura 6.6. Media estricta por usuario

La Figura 6.7 muestra las respuestas correctas obtenidas por cada usuario, independientemente de seleccionar el pasaje correcto (tolerante).

6.3.3 Interpretación de resultados y trabajo futuro

A partir de los resultados obtenidos pueden desprenderse las siguientes conclusiones:

- Los resultados son bajos quizás porque no se ha utilizado ninguna traducción como ayuda. Podría incorporarse una pseudo traducción que realmente ayude a la localización de la respuesta correcta.
- En el caso de los Dominios Relevantes, este método ayuda a encontrar la respuesta correcta pero existen algunos errores debido al escaso contexto de las preguntas. Para subsanar esta escasez

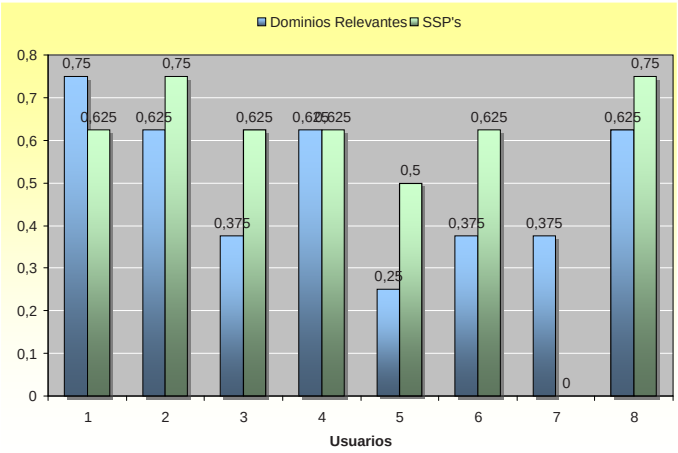


Figura 6.7. Media tolerante por usuario

de información, se podrían añadir más palabras utilizando relaciones de hiperonimia, homonimia, etc.

- En el caso de los SSP's sería necesario mejorar la codificación de los patrones, ya que, actualmente su interpretación resulta complicada, según las encuestas realizadas posteriormente a los usuarios.

Como trabajos futuros se pretende traducir los patrones obtenidos a partir de un método basado en alineamiento de verbos. Además, se pretende mejorar la extracción de pasajes del sistema de recuperación de información mediante la inclusión de información relativa a los Dominios Relevantes.

6.4 Participación en Textual Entailment Recognition

Entre las diferentes tareas de PLN, surge la necesidad de identificar similitudes semánticas entre diferentes fragmentos de texto. Esta tarea recibe el nombre de detección de la implicación textual o Recognising Textual Entailment (RTE).

Nuestra hipótesis es que la utilización de información semántica, podría ser muy útil para resolver el problema de la implicación

textual. Por ello, se ha desarrollado un sistema de detección de la implicación textual participando en las competiciones RTE2 y AVE. Además, directamente relacionada con la detección de la implicación textual se encuentra la detección de la paráfrasis, por ello, también se han realizado varios experimentos que determinan la efectividad de nuestro sistema.

6.4.1 RTE2 PASCAL

La detección de la implicación textual (Dagan et al. (2005)) es una tarea que consiste en determinar dados dos fragmentos de texto, si éstos proporcionan el mismo contenido semántico. Es decir, si a partir de textos diferentes se transmite la misma información. Por ejemplo, dados los textos "Murió debido a la pérdida de sangre" y "Se desangró hasta morir", ambos denotan la misma información, y se podría inferir que el primer texto implica el segundo. Para resolver este problema se han realizado diferentes aproximaciones (Akhmatova (2005), Andreevska et al. (2005), Herrera et al. (2005)), todas ellas evaluadas dentro del marco de PASCAL (Pattern Analysis, Statistical Modelling and Computational Learning) y más concretamente en la tarea de Recognising Textual Entailment Challenge (RTE [13]).

En este tipo de problema, es necesario disponer de cierto conocimiento semántico acerca de los dos contextos implicados. Por ello, nuestro estudio se ha centrado en la determinación de la influencia de la información semántica para la detección de la implicación textual.

La evaluación de las diferentes aproximaciones se ha realizado utilizando el corpus del RTE2 donde los ejemplos proporcionados están balanceados (50 % son verdaderos y 50 % son falsos). Todos los ejemplos han sido extraídos de aplicaciones reales de Extracción de Información, Recuperación de Información, Recuperación de Respuestas y Resúmenes automáticos. En total se proporcionaron 1600 ejemplos, de los cuales, 800 se distribuyeron como corpus de preparación (development data) y los 800 restantes como cor-

[13] http://pascallin.ecs.soton.ac.uk/Challenges/RTE2/

pus de evaluación (test data).

6.4.1.1 Utilización de diferentes corpus para LSA.

Los diferentes experimentos que se han realizado han tenido como objetivo determinar la influencia de la elección de un corpus determinado para establecer la correspondencia semántica entre dos frases. Para ello se han obtenido diferentes espacios semánticos utilizando la técnica de LSA, los cuales se describen a continuación:

- **BNC corpus** *(LSA_BNC_NoTag)*. Resultados utilizando información del corpus BNC con palabras lematizadas.
- **H sentences** *(LSA_LemaH, LSA_NoLemaH)*. Resultados utilizando como corpus las frases H (hipótesis) y construyendo dos matrices diferentes: una con las palabras lematizadas y la otra con las palabras no lematizadas.
- **T sentences***(LSA_LemaT, LSA_NoLemaT)*. Resultados utilizando como corpus las frases T (test) y construyendo dos matrices diferentes: una con las palabras lematizadas y la otra con las palabras no lematizadas.
- **Relevant Domains** *(LSA_RD)*. Resultados utilizando el recurso Dominios Relevantes de cada frase T y de cada frase H.

La medida de evaluación utilizada ha sido la *"Accuracy"* según la Fórmula 6.7:

$$accuracy = \frac{Ejemplos\ \ correctos}{Total\ \ ejemplos} \tag{6.7}$$

En la Tabla 6.22 se muestran los resultados obtenidos para cada uno de los diferentes experimentos realizados.

Para cada experimento se han evaluado los resultados de forma independiente para cada tipo de datos de entrada: IE (Extracción de Información), IR (Recuperación de Información), QA (Búsqueda de respuestas) y SUM (Resumen automático). Como se puede observar, los mejores resultados se han obtenido utilizando las frases de Text (LSA_LemaT) y los Dominios Relevantes (LSA_RD).

Datos	Acc.	IE	IR	QA	SUM
devLSA_BNC_NoTag	49.90	49.87	49.15	50.15	50.43
devLSA_LemaH	53.25	52.00	48.00	54.00	59.00
devLSA_NoLemaH	50.17	50.15	50.03	50.22	50.28
devLSA_LemaT	**56.87**	**51.50**	**58.00**	**56.50**	**61.50**
devLSA_NoLemaT	52.88	50.50	53.00	48.00	60.00
devLSA_RD	**56.98**	**52.25**	**58.60**	**56.83**	**60.25**
testLSA_BNC_NoTag	49.67	49.43	49.00	50.02	50.24
testLSA_LemaH	49.38	52.50	48.50	49.00	47.50
testLSA_NoLemaH	53.37	50.50	54.00	49.00	60.00
testLSA_LemaT	**54.25**	**50.50**	**48.00**	**57.00**	**61.50**
testLSA_NoLemaT	53.63	52.50	50.00	50.00	62.00
testLSA_RD	**54.51**	**50.55**	**48.53**	**56.73**	**62.25**

Tabla 6.22. Resultados usando diferentes corpus y LSA

La primera aproximación utiliza como corpus todas las frases de Text y las frases de Hipótesis como entrada al módulo LSA. En este caso, los resultados son 56,87 % para los datos de preparación y 52,25 % par los datos de evaluación. Estos resultados son mejores que los obtenidos en el experimento LSA_LemaH porque las frases de Text proporcionan mayor información semántica. Por lo tanto, para inferir que dos frases tienen el mismo significado semántico se necesita una aproximación con una base contextual apropiada. La segunda aproximación utiliza como corpus el recurso Dominios Relevantes. En este caso, la matriz contextual inicial se ha obtenido a partir de la información de WordNet Domains. Este espacio semántico se ha utilizado para extraer la similitud entre cada par de frases H-T. Como resultado, se ha obtenido un 56,98 % para los datos de preparación y un 54,51 % para los datos de evaluación. En este caso, los resultados obtenidos han sido bastante buenos debido a que las palabras semánticamente relacionadas comparten las mismas categorías semánticas de WordNet Domains y esta información es muy útil a la hora de relacionar textos relativos a las tares de QA y SUM. En cuanto al resto de experimentos, se demuestra que no se dispone de suficiente información contextual para detectar correctamente la implicación textual.

6.4.1.2 Utilización de la medida del coseno.

Otra serie de experimentos se ha realizado utilizando la medida de similitud del coseno. En este caso, se ha utilizado la medida tradicional de similitud entre documentos y la adaptación de esta medida a los Dominios Relevantes. Los resultados se muestran en la Tabla 6.23.

Datos	Acc.	IE	IR	QA	SUM
devCoseno_DF	52.60	48.63	47.32	55.13	59.32
devCoseno_RD	54.25	50.50	48.00	57.00	61.50
testCoseno_DF	52.18	46.13	49.43	55.34	57.83
testCoseno_RD	54.00	46.50	56.50	56.00	57.00

Tabla 6.23. Results for the cosine measure

A la vista de los resultados queda patente el mejor funcionamiento de la medida del coseno combinada con los Dominios Relevantes. En este caso se alcanza un 54 % tanto para los textos de evaluación como de preparación. Pero estos resultados demuestran que la información contextual dada por las frases no es muy representativa y no proporciona suficiente conocimiento. Por lo tanto, la medida del coseno podría utilizarse combinada con otras fuentes de información.

6.4.1.3 Combinación de LSA y coseno con un sistema de aprendizaje.

Para poder utilizar la información proporcionada por la medida de similitud del coseno y de LSA, éstas se introdujeron como nuevas características en un sistema de aprendizaje automático.

- **MLEnt with previous features** *(MLEnt_Lex, MLEnt_Sem)*. Resultados del sistema previo MLEnt con características léxicas y semánticas.
- **MLEnt with LSA** *(MLEnt_Lex_LSA_LemaT, MLEnt_Sem_LSA_LemaT)*. Resultados del sistema previo MLEnt con LSA. En es-

te caso, se utilizó como corpus para la matriz de LSA las frases de Text con las palabras lematizadas.

- **MLEnt with cosine** *(MLEnt_Lex_cosine,MLEnt_Sem_cosine).* Resultados del sistema previo MLEnt combinado con la medida del coseno. En este caso, el coseno se obtiene utilizando el recurso Dominios Relevantes.
- **MLEnt with LSA and cosine** *(MLEnt_Lex_LSA_LemaT_cosine, MLEnt_Sem_ LSA_LemaT_cosine).* Resultados del sistema previo MLEnt combinando LSA y la medida de similitud del coseno. En este caso, se utilizó LSA con las frases de Text y la medida del coseno con los Dominios Relevantes.

Los resultados obtenidos se muestran en la Tabla 6.24.

Sets	Acc.	IE	IR	QA	SUM
devMLEnt_Lex	56.87	49.50	55.50	51.00	71.50
devMLEnt_Sem	60.12	54.00	61.00	59.00	66.50
devMLEnt_Lex_LSA_LemaT	**62.03**	**56.13**	**62.53**	**60.32**	**69.15**
devMLEnt_Lex_cosine	56.91	49.45	55.62	52.13	70.43
devMLEnt_Lex_LSA_LemaT_cosine	57.13	49.50	55.50	52.50	71.00
devMLEnt_Sem_LSA_LemaT	**62.56**	**57.13**	**62.83**	**60.54**	**69.75**
devMLEnt_Sem_cosine	60.21	54.13	61.06	59.14	66.54
devMLEnt_Sem_LSA_LemaT_cos	**61.75**	**56.00**	**59.50**	**62.50**	**69.00**
testMLEnt_Lex	51.75	52.00	53.50	55.50	46.00
testMLEnt_Sem	54.25	50.00	55.50	47.50	64.00
testMLEnt_Lex_LSA_LemaT	**55.01**	**51.23**	**55.83**	**47.96**	**65.03**
testMLEnt_Lex_cosine	52.57	49.50	44.95	53.73	62.13
testMLEnt_Lex_LSA_LemaT_cosine	54.87	46.50	53.00	56.00	64.00
testMLEnt_Sem_LSA_LemaT	**56.18**	**52.03**	**56.53**	**50.14**	**66.03**
testMLEnt_Sem_cosine	54.42	50.22	55.62	47.61	64.25
testMLEnt_Sem_LSA_LemaT_cos	**56.50**	**53.00**	**58.00**	**57.50**	**57.50**

Tabla 6.24. Resultados para la combinación de MLEnt con LSA y el coseno

Como muestra la Tabla 6.24, los experimentos que se han llevado a cabo combinando los valores de LSA y el coseno como nuevas características, mejoran los resultados previos del sistema MLEnt. Por tanto, podemos concluir que añadir información semántica a un sistema de aprendizaje automático proporciona mayor efectividad. De hecho, el mejor resultado obtenido es de

un 62 % para el conjunto de datos de preparación y de un 57 % para el conjunto de datos de evaluación. Estos resultados se han obtenido a partir del experimento realizado tras la combinación de LSA con el coseno y el sistema MLEnt.

6.4.1.4 Comparativa con otros sistemas participantes.

En la competición para la detección de la implicación textual RTE2, participaron 23 equipos. Cada equipo podía enviar hasta dos ejecuciones para evaluar sus resultados. La Tabla 6.25 presenta los resultados obtenidos en términos de la medida de *"Accuracy"*.

Desde el punto de vista semántico, los resultados obtenidos ofrecen una mejora con respecto al sistema inicial MLEnt. De esta forma, se puede afirmar que la combinación del sistema de aprendizaje MLEnt con recursos semánticos tales como, los dominios relevantes y la semántica latente pone de manifiesto la utilidad de añadir información semántica para la detección de la implicación textual.

Además, analizando los resultados obtenidos con respecto al resto de participantes, la media obtenida ronda el 58 %, por lo que nuestro sistema obtiene resultados significativos.

6.4.2 AVE CLEF2006

En esta competición se participó en la tarea Answer Validation Exercise (AVE) (Peñas et al. (2006a)), con el sistema basado en aprendizaje MLEnt desarrollado en (Kozareva y Montoyo (2006), Kozareva et al. (2006)). Este sistema fue utilizado para la tarea RTE y fue posteriormente adaptado para la tarea AVE.

En la competición AVE el objetivo es determinar si a partir de un *snippet* proporcionado por un sistema de búsqueda de respuestas (T-Texto) se puede extraer la información proporcionada en un texto (H-Hipótesis). De esta forma, se podría evaluar de forma automática el funcionamiento de sistemas de Question Answering.

En la edición de 2006 AVE se planteó como tarea multilingüe para detectar la corrección de las respuestas dadas por un siste-

Sistema	Acc
Adams (Dallas)	0.6262
Bos (Rome & Leeds)	0.6162
	0.6062
Burchardt (Saarland)	0.5900
	0.5775
Clarke (Sussex)	0.5275
	0.5475
De Marneffe (Stanford)	0.5763
	0.6050
Delmonte (Venice)	0.5475
Ferrández (Alicante)	0.5563
	0.5475
Herrera (UNED)	0.5975
	0.5887
Hickl (LCC)	0.7538
Inkpen (Ottawa)	0.5800
	0.5825
Katrenko (Amsterdam)	0.5900
	0.5713
Kouylekov (ITC-irst & Trento)	0.5725
	0.6050
Combinación MLEnt – Lex – LSA	0.5501
Combinación MLEnt – Sem - LSA	0.5618
Combinación MLEnt – Sem - LSA – coseno	0.5650
Kozareva (Alicante)	0.5487
	0.5500
Litkowski (CL Research)	0.5813
	0.5663
Marsi (Tilburg & Twente)	0.6050
Newman (Dublin)	0.5250
	0.5437
Nicholson (Melbourne)	0.5288
	0.5088
Nielsen (Colorado)	0.5962
	0.5875
Rus (Memphis)	0.5900
	0.5837
Schilder (THomson & MInnesota)	0.5437
	0.5550
Tatu (LCC)	0.7375
Vanderwende (Microsoft Research & Stanford)	0.6025
	0.5850
Zanzotto (MIllan & Rome)	0.6388
	0.6250

Tabla 6.25. Evaluación de sistemas en RTE2

Universitat d'Alacant
Universidad de Alicante

ma de QA en diferentes idiomas: español, inglés, alemán, francés, italiano, holandés y portugués.

6.4.2.1 Módulo de solapamiento de palabras: Sistema MLEnt.

El sistema MLEnt está compuesto de dos módulos: módulo de palabras superpuestas y módulo de similitud semántica. Dado que la tarea AVE es una tarea multilingüe, el módulo de similitud semántica no se ha utilizado debido a que éste utiliza la información proporcionada por WordNet y es muy costoso adaptar este módulo para diferentes idiomas.

Los atributos utilizados por el módulo de palabras superpuestas son los siguientes:

- **n-gramas.** Busca posiciones comunes de unigramas entre el Texto y la Hipótesis. De acuerdo a este atributo el par Texto-Hipótesis será correcto si ambos comparten las mismas palabras. De la misma forma, este atributo determina que un par no es correcto si no contienen ninguna palabra en común. Este atributo no considera información de similitud semántica, así por ejemplo si "vehículo" y "coche" aparecen respectivamente en T y H son consideradas palabras sin ningún tipo de relación y por tanto, totalmente distintas. Otro punto débil de este atributo es que no tiene en cuenta el orden de las palabras y la estructura de las frases. En este caso, frases del tipo "Mary calls the police" y " The police calls Mary", contienen las mismas palabras, pero con este atributo el resultado sería que no infieren el mismo significado. Para solventar este problema se han creado los siguientes atributos LongCS y skip-gramas.
- **LongCS (Longest Common Subsequence).** Obtiene secuencias de palabras no consecutivas de cualquier longitud, entre el Texto y la Hipótesis. Un valor elevado de LongCS significa sentencias similares. El valor de LongCS entre cada par T(m)-H(n), donde m es la longitud del Texto y n es la longitud de la hipótesis, se determina como $\frac{LongCS(T,H)}{n}$.

- **skip-gramas.** Representa cualquier par de palabras en una oración con un número indeterminado de palabras entre ellas. Una vez determinados todos los pares de palabras del Texto y la Hipótesis, se realiza el conteo de los skip-gramas de la siguiente forma $\frac{skip\_gramas(T,H)}{C(n,numero\_skip-gramas)}$. Donde $skip-gramas(T,H)$ hace referencia al número de skip-gramas comunes entre T y H, $C(n, numero\_skip-gramas)$ es una función combinatoria, donde n es el número de palabras en H y numero_skip-gramas se corresponde con el número de n-gramas comunes entre T y H. De acuerdo a este atributo el par T-H será correcto cuantos más skip-gramas tengan en común. Por ejemplo, para las siguientes oraciones:
  S1: "Mary calls the police"
  S2: "Mary called the police"
  S3: "The police called Mary"
  Mediante el atributo skip-gramas se deduce que las frases S1 y S2 tienen una relación de similitud más fuerte que S1 y S3 o S2 y S3. Sin embargo, los atributos n-gramas y LongCS no son tan efectivos y no pueden determinar la similitud correctamente.
- **mapeo_numérico.** Se identifican los números presentes en T y H y se verifican. Para frases donde no existen números, este atributo asigna en valor NO para el par. De acuerdo a este atributo, el par T-H será correcto cuanto los números de T y H coincidan.

El conjunto de atributos descrito ha sido evaluado únicamente sobre inglés y español, debido a que sólo se proporcionó corpus de entrenamiento para estos dos idiomas. Para la fase de entrenamiento se utilizaron los clasificadores SVM y kNN, junto con la observación de la medida IG (Information Gain), para los dos idiomas y diferentes tamaños de corpus de entrenamiento. IG es una medida que indica a partir de un conjunto de características cuáles son las más importantes. De acuerdo a IG, los dos atributos que proporcionan mejores resultados son LongCS y skip-gramas. Para la característica de solapamiento de palabras, el sistema genera dos salidas, una obtenida a partir del atributo LongCS y otra obtenida por el atributo de skip-gramas.

Universitat d'Alacant

Para el resto de idiomas a los que no se proporcionó corpus de entrenamiento se tuvo que ajustar los atributos LongCS y skip-gramas. Dado que los atributos utilizados dependen de la longitud del solapamiento de palabras normalizado por el número total de palabras presentes en H, fue posible adaptar estos atributos. Así, utilizando las medidas de desviación estándar obtenidas para LongCS y skip-gramas en español e inglés, se adaptaron al resto de idiomas.

6.4.2.2 Módulo de similitud semántica: LSA.

Una de las características de LSA es la de detectar similitudes semánticas entre textos que aun no compartiendo las mismas palabras, puedan estar relacionados. Esta capacidad ya fue comentada en el capítulo anterior con más detalle.

En nuestro caso, para poder aplicar LSA sobre la tarea AVE se utilizó como corpus para construir la matriz conceptual las frases del Texto. Esta decisión fue tomada debido al estudio realizado en (Vázquez et al. (2006)), donde al utilizar las frases de T como corpus, la evaluación sobre la tarea de RTE obtenía mejores resultados. Por tanto, para cada uno de los distintos idiomas - inglés, español, italiano, alemán, holandés, portugués y francés - se construyeron diferentes matrices conceptuales utilizando las oraciones del Texto del corpus proporcionado por AVE.

A partir de las matrices conceptuales obtenidas en cada caso, se pueden establecer relaciones de similitud entre términos, frases o documentos. En nuestros experimentos, dado que el objetivo final era determinar si dos frases T-H tenían relación semántica, se utilizó la similitud entre frases. El resultado tras aplicar cada frase de H sobre la matriz conceptual, es un listado ordenado de mayor a menor similitud con las diferentes frases de T.

En los siguientes ejemplos se muestran los datos de entrada. Para cada pregunta Q se proporcionaba un pasaje de texto T del cual se debía inferir H. Los ejemplos muestran una instancia para la que se debe devolver FALSO (ejemplo 1) y otra para la que se debe devolver VERDADERO (ejemplo2).

Ejemplo 1:

<*pair* id="4525" value="NO" task="QA">
<*q* lang="EN" src="clef2006" type="OBJECT"> **What is Atlantis** </*q*>
<*t* doc="096222"> **TO ATLANTIS' CREW. From Associated Press NASA briefly lost contact with the space shuttle Atlantis and its six astronauts Sunday because of crossed radio signals. The problem occurred as Atlantis switched from one Tracking and Data Relay Satellite to another, a routine procedure during Atlantis nor its crew was in any danger, and no science data was lost, said Mission Control with Atlantis was restored after eight minutes, but it was an hour before engineers realized crossed signals,**</*t*>
<*h*>**Atlantis is ATLANTIS THE LOST EMPIRE.**</*h*>
</*pair*>

Este es un par T-H extraído de la colección de test de AVE, para la cual el sistema debe devolver "NO" como respuesta. En este caso, a partir del resultado obtenido por un sistema de QA para la pregunta *"What is Atlantis?"*, se trata de establecer si a partir del texto T se puede inferir la respuesta H. El resultado tras aplicar el Módulo LSA es el valor 0,402886, se considera por tanto, que no existe implicación entre ambos pares.

Ejemplo 2:

<*pair* id="7818" value="YES" task="QA">
<*q* lang="EN" src="clef2006" type="OBJECT"> **What is Atlantis** </*q*>
<*t* doc="LA110794-0104"> **NASA briefly lost contact with the space shuttle Atlantis and its six astronauts Sunday because of crossed radio signals.**</*t*>
<*h*> **Atlantis is the space shuttle.**</*h*>
</*pair*>

En este ejemplo, el resultado debe ser "YES". Para la pregunta *"What is Atlantis?"*, y con el par T-H, sí se puede inferir el contenido de H a partir de T. En este caso, el módulo LSA obtiene el valor 0,905481, por lo que sí existe implicación para el par.

6.4.2.3 Módulo combinatorio.

El último paso para la obtención del sistema final presentado en AVE es realizar la combinación del resultado de los módulos de las secciones anteriores: módulo de solapamiento de palabras y módulo de LSA. La combinación de ambos módulos se realiza mediante una estrategia de votación.

Para garantizar una buena elección en la votación se realizaron diferentes pruebas para comprobar la compatibilidad de los distintos módulos. La medida utilizada fue el coeficiente Kappa (Cohen (1960), Pedersen (2002)), que permite establecer el grado de acuerdo entre los distintos clasificadores según la Fórmula 6.8.

$$Kappa = \frac{P_0 - P_e}{1 - P_e} \tag{6.8}$$

Siendo $P_0 = \frac{num\ acuerdos}{num\ acuerdos + num\ desacuerdos}$ y $P_e = \sum_{i=1}^{n} (P_{i1} \times P_{i2})$

Donde:

n = número de categorías

i = número de la categoría (de 1 hasta n)

P_{i1} = proporción de ocurrencia de la categoría i para el observador 1.

P_{i2} = proporción de ocurrencia de la categoría i para el observador 2.

Un valor de Kappa elevado indica un alto grado de acuerdo entre los clasificadores por lo que no existe una mejora aparente tras aplicar la votación. En cambio, un valor de Kappa bajo indica un grado de acuerdo muy bajo por lo que se podrá apreciar una mejora tras la combinación de los resultados. Landis y Koch (Landis y Koch (1977)) propusieron unos márgenes para valorar el grado de acuerdo en función del índice Kappa Tabla 6.26.

Para cada par T-H de AVE, se obtuvieron diferentes resultados usando LongCS, skip-gramas y LSA. La medida Kappa se utilizó evaluando los tres resultados juntos y también se evaluó por pares de resultados. Los experimentos desarrollados para inglés

Universitat d'Alacant
Universidad de Alicante

Kappa	Grado de acuerdo
< 0	Sin acuerdo
$0 - 0{,}2$	Insignificante
$0{,}2 - 0{,}4$	Bajo
$0{,}4 - 0{,}6$	Moderado
$0{,}6 - 0{,}8$	Bueno
$0{,}8 - 1$	Muy bueno

Tabla 6.26. Grado de acuerdo Kappa

y español (de los únicos idiomas de los que se disponía de corpus de entrenamiento), demostraron que la mejor combinación era LongCS con skip-gramas y LongCS, skip-gramas y LSA. Por tanto, se presentaron dos ejecuciones distintas.

Una vez la medida Kappa determinó las salidas que debían ser combinadas, se aplicó la técnica de votación. Mediante esta técnica, se combinaron las distintas salidas en una única predicción. Las salidas generadas para LongCS, skip-gramas y LSA fueron evaluadas y se escogió la respuesta con mayor número de votos. Para LongCS y skip-gramas, no se pudo aplicar votación dado que sólo hay dos clasificadores. En este caso, se tomó como estrategia que si no había acuerdo entre los dos clasificadores se respondiera "NO", y si había consenso se respondiera lo que indicaran los clasificadores.

Universitat d'Alacant
Universidad de Alicante

6.4.2.4 Evaluación de resultados.

La evaluación de los resultados obtenidos se realizó sobre los diferentes idiomas de la tarea AVE: inglés, español, alemán, francés, italiano, holandés y portugués. En la Tabla 6.27 se muestran los resultados para las distintas ejecuciones individuales del módulo de solapamiento de palabras y LSA, así como los resultados obtenidos para las dos combinaciones LongCS y skip-gramas y LongCS, skip-gramas y LSA.

Para la evaluación de los resultados obtenidos se han utilizado las siguientes medidas de evaluación:

$$precision = \frac{\#contestados\ correctamente\ como\ YES}{\#total\ contestados\ como\ YES} \quad (6.9)$$

$$recall = \frac{\#contestados\ correctamente\ como\ YES}{\#total\ pares\ YES} \quad (6.10)$$

$$F\text{-}score = \frac{2 * recall * precision}{recall + precision} \quad (6.11)$$

Estas medidas fueron proporcionadas por los organizadores de la tarea AVE. De acuerdo a un estudio realizado en (Peñas et al. (2006b)), el 25 % de los pares son ciertos y el 75 % son falsos. Por tanto, el funcionamiento de los sistemas presentados debería evaluarse teniendo en cuenta los pares etiquetados como "YES".

Para los diferentes idiomas desarrollamos a continuación una pequeña descripción de las tareas realizadas:

- **Inglés.** Para este idioma, se dispuso de una fase de entrenamiento utilizando los datos del corpus ENGARTE[14] proporcionado. Los resultados obtenidos en este experimento se utilizaron como indicadores para seleccionar los mejores atributos del conjunto inicial. El mejor atributo del módulo de solapamiento de palabras fue LongCS tanto para el test como para el corpus de entrenamiento. Esto demuestra que un tercio de de los pares del AVE pueden ser resueltos correctamente, simplemente considerando las secuencias solapadas entre dos textos.
  Los atributos skip-gramas y LSA obtienen alrededor de un 27 % de precisión. La combinación de ambos no supuso ninguna mejora en el corpus de test pero sí se incrementó en un 2 % en el corpus de entrenamiento. El mejor resultado para este idioma se obtuvo mediante la combinación de LongCS, skip-gramas y LSA. Esto demuestra que el atributo LSA detecta correctamente pares que los otros dos atributos no son capaces de clasificar. De acuerdo con la medida estadística z' con un nivel de confianza de 0.975 el incremento es significativo.

[14] `http://nlp.uned.es/QA/ave`

Idioma	Precision	Recall	F-score
Inglés_LongCS	15,22	80,93	28,57
Inglés_Skip	16,91	69,30	27,18
Inglés_LSA	23,29	30,23	26,31
Inglés_LongCS&Skip	18,33	64,65	28,56
Inglés_LongCS&Skip&LSA	24,92	69,77	**36.72**
Español_LongCS	44,21	66,62	**53.15**
Español_Skip	37,24	43,07	39,94
Español_LSA	34,15	14,45	20,31
Español_LongCS&Skip	47,48	39,34	43,03
Español_LongCS&Skip&LSA	40,65	76,15	**53.01**
Alemán_LongCS	38,90	60,56	**47.37**
Alemán_Skip	34,37	43,91	38,55
Alemán_LSA	11,43	1,13	2,06
Alemán_LongCS&Skip	41,30	37,68	39,41
Alemán_LongCS&Skip&LSA	36,34	67,42	**47.22**
Francés_LongCS	33,96	67,09	**45.09**
Francés_Skip	30,48	46,38	36,78
Francés_LSA	32,36	15,88	21,31
Francés_LongCS&Skip	38,36	43,69	40,85
Francés_LongCS&Skip&LSA	34,44	73,62	**46.93**
Italiano_LongCS	25,78	70,59	**37.77**
Italiano_Skip	21,96	86,10	34,99
Italiano_LSA	29,16	22,45	25,37
Italiano_LongCS&Skip	21,64	88,77	34,80
Italiano_LongCS&Skip&LSA	28,30	72,19	**40.66**
Holandés_LongCS	14,26	90,12	24,62
Holandés_Skip	15,80	67,901	**25.64**
Holandés_LSA	13,88	12,34	13,07
Holandés_LongCS&Skip	18,90	67,90	**29.57**
Holandés_LongCS&Skip&LSA	14,84	90,12	25,48
Portugués_LongCS	12,50	3,90	5,94
Portugués_Skip	8,00	21,00	11,58
Portugués_LSA	11,26	12,76	11,96
Portugués_LongCS&Skip	19,04	12,77	**15.29**
Portugués_LongCS&Skip&LSA	19,15	14,76	**16.67**

Tabla 6.27. Resultados para la evaluación de AVE

- **Español.** Para español se desarrolló otra fase de entrenamiento utilizando como corpus SPARTE. Para el corpus del test los mejores resultados se obtuvieron con el atributo LongCS llegando a un valor aproximado de 53 %. El resultado tras la aplicación de la votación sobre la combinación de los tres atributos obtuvo el mismo valor que el atributo LongCS por separado. Esto es debido en gran parte a la baja cobertura de LSA, que depende del número y tipo de palabras de las frases de Texto.
- **Alemán, francés e italiano.** Para estos tres idiomas los mejores resultados se obtuvieron con el atributo LongCS y la combinación de los tres atributos. El rango de valores proporcionados por la medida F-score se sitúa entre 40 % a 47 %. Como se puede observar, los resultados de LSA son más bajos respecto a los obtenidos con los atributos del módulo de solapamiento. Esto es debido a que el grado de similitud de 0.8 sobre el que se propuso determinar si un par T-H era correcto, depende del tipo de palabras contenidas en T y debe ser estudiado con detalle para cada idioma.
- **Holandés y portugués.** Para estos dos idiomas se obtuvieron los peores resultados. Cabe destacar que en el caso del holandés el atributo skip-gramas obtiene mejores resultados que LongCS. Este hecho puede estar relacionado con el origen de este idioma y el orden existente entre las palabras, ya que, los skip-gramas buscan posiciones independientes unas de otras a diferencia de los n-gramas que buscan posiciones contiguas. Para portugués, LSA obtiene mejores resultados que cualquiera de los atributos del módulo de solapamiento de palabras. El resultado tras la votación para portugués obtuvo un 4 % de mejora frente a los clasificadores individuales.

A la vista de los resultados obtenidos tras la evaluación de los distintos atributos o clasificadores por separado o mediante votación, podemos concluir que éstos funcionan correctamente independientemente del idioma utilizado. Además, la estrategia de votación mejora en la mayoría de los casos los resultados obtenidos por los clasificadores de forma individual.

6.4.2.5 Comparativa con otros sistemas participantes.

En la competición AVE participaron 11 equipos diferentes, los cuales, realizaron la evaluación de sus sistemas en diferentes idiomas. En nuestro caso, nuestro sistema se ha adaptado de tal forma que es capaz de trabajar sobre diferentes idiomas, tal y como se ha mostrado en las secciones previas. Sin embargo, debido a las características específicas de cada idioma y a la forma establecer conexiones entre palabras, la efectividad del sistema se ve en algunos casos truncada.

La Tabla 6.28 muestra los resultados obtenidos por los diferentes sistemas, junto con la posición alcanzada por el sistema definido en las secciones anteriores.

A la vista de los resultados obtenidos, podemos concluir que el sistema propuesto, resultado de la combinación de recursos de índole semántica con un sistema de aprendizaje automático, obtiene buenos resultados. Cabe destacar que este sistema se ha aplicado a todos los idiomas de la tarea, adaptando en cada caso únicamente el módulo de LSA. Esta adaptación simplemente requiere que la codificación de la matriz conceptual se realice a partir de la información de los textos de cada idioma respectivamente.

6.4.3 Detección de paráfrasis

Estrechamente relacionada con el concepto de Implicación Textual se encuentra la paráfrasis. Mediante la paráfrasis se puede reescribir un texto utilizando sinónimos (paráfrasis mecánica) o cambiando la estructura, el contenido ... (paráfrasis constructiva), pero siempre conservando el significado original. Por ejemplo, "vehículo" y "coche", "X está casado con Y" y "X es el marido de Y", etc.

Existen diferentes aproximaciones que tratan de identificar la paráfrasis entre textos. Muchas de ellas se centran en la extracción de reglas que detectan la paráfrasis (Lin y Pantel (2001), Barzilay y McKeown (2001), Barzilay y McKeown (2003)). Otras identifican la paráfrasis entre textos a partir del solapamiento de

Sistema	Prec	Rec	F
Inglés			
COGEX	0.3261	0.7576	0.4559
ZNZ -TV 2	0.2838	0.7424	0.4106
itc-irst	0.3090	0.5354	0.3919
ZNZ -TV 1	0.2707	0.6263	0.3780
UA_comb	0.2492	0.6977	0.3672
uaofe 2	0.2040	0.7172	0.3177
uaofe 1	0.2144	0.5404	0.3070
utwente.ta	0.3313	0.2778	0.3022
utwente.lcs	0.2692	0.2828	0.2759
ebisbal	0.2143	0.0455	0.075
Español			
COGEX	0.527	0.7139	0.6063
UNED 1	0.467	0.7168	0.5655
UNED 2	0.4652	0.7079	0.5615
NED	0.4364	0.6796	0.5315
UA_comb	0.4065	0.7615	0.5301
R2D2	0.4387	0.5648	0.4938
utwente.ta	0.4811	0.4560	0.4682
utwente.lcs	0.5507	0.3562	0.4326
Alemán			
FUH 1	0.5839	0.5058	0.5420
FUH 2	0.7293	0.3837	0.5029
UA_comb	0.3634	0.6742	0.4722
utwente.lcs	0.4	0.0872	0.1432
Francés			
UA_comb	0.3444	0.7362	0.46.93
LIRAVE	0.4327	0.0638	0.1112
utwente.lcs	0.4625	0.0525	0.0943
Italiano			
UA_comb	0.2830	0.7219	0.4066
utwente.lcs	0.3281	0.1123	0.1673
Holandés			
utwente.ta	0.2874	0.5926	0.3871
UA_comb	0.1890	0.6790	0.2957
utwente.lcs	0.2	0.2469	0.2201
Portugués			
utwente.lcs	0.5783	0.2553	0.3542
UA_comb	0.1915	0.1476	0.1667

Tabla 6.28. Evaluación sistemas participantes en AVE

Universitat d'Alacant
Universidad de Alicante

palabras o de la similitud entre palabras. El problema de estas aproximaciones es que representan de forma global los conceptos y por tanto, no reflejan realmente el significado de los contextos.

Nuestra propuesta para detectar la paráfrasis utiliza como fuente de información el recurso léxico Dominios Relevantes. De esta forma, se pueden establecer las diferentes relaciones semánticas entre diferente segmentos de texto. La hipótesis en la que se basa esta aproximación es que las etiquetas semánticas o dominios determinan la coherencia de los textos, ya que, palabras relacionadas semánticamente comparten dominios similares y maximizan la similitud entre textos. Para establecer las similitudes entre los diferentes segmentos de texto se utiliza la técnica de LSA, cuya matriz se obtiene tomando como contextos las palabras de las glosas de WordNet asociadas a cada uno de los diferentes dominios.

6.4.3.1 Utilización de WordNet Domains y SUMO.

Además de WND existe otra ontología más general construida sobre las definiciones de WordNet, estamos hablando de la ontología SUMO. Esta ontología al igual que ocurría con WND extiende las relaciones entre palabras utilizando categorías semánticas o dominios. El objetivo de este estudio es determinar la influencia ejercida por el tipo de ontología utilizada para establecer las relaciones semánticas entre diferentes contextos.

En la Figura 6.8 se muestra una parte de cada una de las jerarquías de WND y SUMO. Como se puede apreciar los conceptos representados en la jerarquía de SUMO son mucho más genéricos que los representados en WND.

La evaluación de este método se ha realizado a partir de la obtención de dos matrices conceptuales distintas: una matriz obtenida a partir de WND y otra obtenida a partir de SUMO.

El proceso de obtención de las matrices es básicamente el mismo. El primer paso es obtener los Dominios Relevantes de cada palabra utilizando la jerarquía de WND y la jerarquía de SUMO. La obtención de los dominios relevantes se realiza a partir de la fórmula del Ratio de Asociación, que determina la relevancia de un dominio con respecto a una palabra. Con esta información se

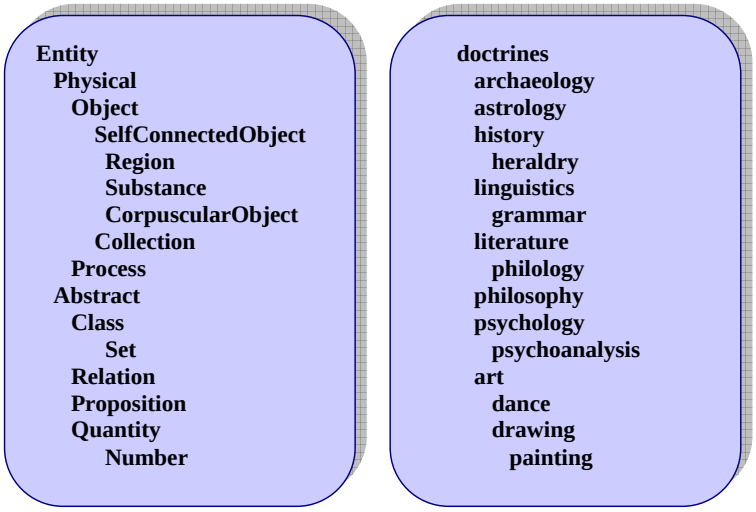


Figura 6.8. Comparación de las jerarquías SUMO y WND

construye la matriz conceptual, utilizando en lugar de documentos como columnas, las etiquetas semánticas de los dominios y como valor en cada celda, el valor del Ratio de asociación de cada palabra con respecto a cada dominio.

Una vez obtenidas las matrices conceptuales se realiza su descomposición en valores singulares, reduciendo ambas matrices a 100 dimensiones.

Universitat d'Alacant
Universidad de Alicante

6.4.3.2 Ejemplo ilustrativo.

Para ilustrar la aplicación de los Dominios Relevantes y LSA se muestra a continuación un ejemplo de resolución de paráfrasis. En este caso, dados dos fragmentos de texto, se van a extraer sus correspondientes espacios conceptuales y se va a determinar un valor de similitud entre ambos. El valor a partir del cual dos textos son considerados como paráfrasis uno del otro se ha obtenido empíricamente a partir de una serie de experimentos previos sobre un corpus de entrenamiento.

El primer paso, es obtener los lemas de las palabras de los textos implicados (Figura 6.9). Para ello se ha utilizado el analizar sintáctico Tree-Tagger. Se ha optado por extraer los lemas debido a que la matriz conceptual de LSA se ha construido a partir de las palabras lematizadas de WordNet. En este ejemplo, se han extraído los dominios relevantes de nombres, verbos, adjetivos y adverbios, para ambos segmentos. Las palabras que aparecen subrayadas en la Figura 6.9, son aquellas para las que se han considerado sus dominios relevantes.

Text Segment 1: Women who eat potatoes and other tuberous vegetables during pregnancy may be at risk of triggering type 1 diabetes in their children, Melbourne researchers believe.

Text Segment 2: Australian researchers believe they have found a trigger of type 1 diabetes in children - their mothers eating potatoes and other tuberous vegetables during pregnancy.

Figura 6.9. Textos número 1634 del corpus

La Figura 6.10 muestra los dominios relevantes de cada una de las palabras de los segmentos según la medida del Ratio de Asociación.

Una vez determinados los dominios relevantes de los dos segmentos de texto, es necesario determinar el grado de similitud entre ambos. Para ello, se utiliza la técnica de LSA que obtiene los dominios que tienen en común ambos segmentos de texto (ver Figura 6.29). Con esta información se seleccionan los dominos con la probabilidad más elevada que coincidan con los dominios relevantes extraídos anteriormente.

Finalmente, se obtiene el valor del dominio más apropriado de acuerdo a los valores de similitud obtenidos. En este caso, el dominio seleccionado es APPLIED_SCIENCE.

6.4.3.3 Evaluación.

Text Segment 1:
woman={sexuality 0.236904, fashion 0.074808, person 0.072525, athletics 0.048517, jewellery 0.042176}
eat={gastronomy 0.168685, ecology 0.034430, folklore 0.026185, physiology 0.017776, anthropology 0.012501}
potato={agriculture 0.056402, gastronomy 0.009348, entomology 0.004056, racing 0.003743, medicine 0.002409}
tuberous={agriculture 0.000782, biology 0.000284, botany 0.003115, botany 0.003115, gastronomy 0.002218}
vegetable={gastronomy 0.040430, zootechnics 0.023290, agriculture 0.022609, earth 0.009891, body_care 0.009335}
pregnancy={surgery 0.027848, physiology 0.025092, medicine 0.005344, anatomy 0.002291, color 0.001075}
risk={insurance 0.049295, exchange 0.015876, enterprise 0.013756, industry 0.001393, commerce 0.001289}
trigger={commerce 0.002437, computer_science 0.001999, factotum 0.000088 }
type={zoology 0.052495, philology 0.048450, bowling 0.043687, publishing 0.023217, biology 0.018311}
diabetes={pharmacy 0.006108, medicine 0.005782, alimentation 0.000724, time_period 0.000290, factotum 0.000020...}
child={ethnology 0.008168, acoustics 0.006704, color 0.002306, body_care 0.001732, economy 0.001036}
researcher={person 0.000636, factotum 0.000010}
believe={doctrines 0.195175, theology 0.155574, pure_science 0.137293, folklore 0.079765, religion 0.067227}
Text Segment 2:
researcher={person 0.000636, factotum 0.000010}
believe={doctrines 0.195175, theology 0.155574, pure_science 0.137293, folklore 0.079765, religion 0.067227}
find={zoology 0.102364, chemistry 0.072100, statistics 0.045846, geology 0.043141, astrology 0.042836}
trigger={commerce 0.002437, computer_science 0.001999, factotum 0.000088 }
type={zoology 0.052495, philology 0.048450, bowling 0.043687, publishing 0.023217, biology 0.018311}
diabetes={pharmacy 0.006108, medicine 0.005782, alimentation 0.000724, time_period 0.000290, factotum 0.000020}
child={ethnology 0.008168, acoustics 0.006704, color 0.002306, body_care 0.001732, economy 0.001036}
mother={archaeology 0.014541, anthropology 0.003027, computer_science 0.000241, administration 0.000241, biology 0.000239}
eat={gastronomy 0.168685, ecology 0.034430, folklore 0.026185, physiology 0.017776, anthropology 0.012501}
potato={agriculture 0.056402, gastronomy 0.009348, entomology 0.004056, racing 0.003743, medicine 0.002409}
tuberous={agriculture 0.000782, biology 0.000284, botany 0.003115, botany 0.003115, gastronomy 0.002218}
vegetable={gastronomy 0.040430, zootechnics 0.023290, agriculture 0.022609, earth 0.009891, body_care 0.009335}
pregnancy={surgery 0.027848, physiology 0.025092, medicine 0.005344, anatomy 0.002291, color 0.001075}

Figura 6.10. Los cinco primeros dominios relevantes de cada palabra

LSA dominios en segmento 1		LSA dominios en segmento 2	
Dominio	**Similitud**	**Dominio**	**Similitud**
applied_science	0.770537	applied_science	0.793825
pharmacy	0.740445	pharmacy	0.777943
philology	0.717400	ecology	0.713885
publishing	0.716576	transport	0.709478
theology	0.714463	biology	0.705481
pedagogy	0.705165	botany	0.701570
telecommunication	0.700763	university	0.694129
university	0.698827	publishing	0.693940
psychoanalysis	0.697876	chemistry	0.693747

Tabla 6.29. LSA listado con los nuevo dominios relevantes para cada texto

Para evaluar la corrección de los resultados se han realizado diversos experimentos sobre un corpus de paráfrasis [15]. El proceso de evaluación consiste en determinar dados dos segmentos de texto si existe paráfrasis entre ambos.

El corpus utilizado (Dolan et al. (2004)) ha sido extraído de la web. El número de instancias de entrenamiento es de 4076 y el número de instancias de test es de 1725.

Un ejemplo de segmentos de texto es: *"Inhibited children tend to be timid with new people, objects, and situations, while uninhibited children spontaneously approach them."* y *"Simply put, shy individuals tend to be more timid with new people and situations."*

Las medidas de evaluación utilizadas han sido: precisión, recall, accuracy y f-measure.

Se han desarrollado dos tipos de experimentos. El primero, estudia cómo representar el concepto subyacente a dos segmentos de texto, utilizando WND y SUMO. En el segundo, se estudia si el uso de ontologías más genéricas, produce mejores resultados. Los resultados obtenidos se muestran en la Tabla 6.30. En esta tabla se muestran los resultados obtenidos en el corpus de entrenamiento y en el corpus de test. Así como también una serie de umbrales utilizados, con los que se concluye que un umbral de 0.4 es el más adecuado.

[15] http://research.microsoft.com/en-us/projects/paraphrase/default.aspx

	Datos	Umb	Acc	Prec	Rec	F
WordNet Domains	Train	0.8	80.29	72.97	70.83	71.89
		0.6	97.35	68.91	96.07	80.26
		0.4	**98.52**	68.36	97.82	**80.48**
	Test	0.8	80.34	72.08	70.44	71.25
		0.6	97.10	67.50	95.64	79.14
		0.4	**98.26**	66.84	97.38	**79.27**
SUMO	Train	0.8	38.59	81.69	09.08	16.34
		0.6	94.28	69.44	91.53	78.97
		0.4	**96.27**	68.93	94.47	**79.71**
	Test	0.8	40.05	81.29	09.85	17.57
		0.6	93.50	68.67	90.23	77.99
		0.4	**95.18**	68.11	92.76	**78.55**
Text similarity approach	Test	–	68.80	74.10	81.70	77.70

Tabla 6.30. Representación conceptual para identificar la paráfrasis

Para la resolución de la paráfrasis esta aproximación proporciona no sólo el significado del texto sino también su concepto semántico global. Durante el proceso de entrenamiento y de test WND y SUMO obtienen resultados similares. Sin embargo, WND proporciona resultados más precisos. Las diferencias más notables entre los dos experimentos se muestran con umbrales de valores elevados. Mientras que WND varía alrededor de un 10 % entre los diferentes umbrales, SUMO varía del 16 al 79 %. Esto debido en gran parte a la jerarquía utilizada en ambas ontologías. En WND los dominios se pueden solapar con una alta probabilidad mientras que en la jerarquía genérica de SUMO este solapamiento es menos evidente.

En la Tabla 6.30 también se muestra una comparativa con los resultados obtenidos por la aproximación mediante similitudes (Corley y Mihalcea (2005)). Se puede apreciar que nuestra aproximación obtiene mejores resultados. Ya que, el establecimiento de similitudes palabra−palabra o texto−texto, no determinan exactamente el significado del texto. En nuestro caso, se determina la similitud entre palabras pertenecientes a distintas categorías sintácticas en base al concepto semántico subyacente.

Algunas limitaciones con respecto a la representación conceptual de los segmentos de texto son debidas a la inclusión del dominio FACTOTUM, cuando no se puede clasificar alguna palabra.

Con respecto a la ontología SUMO se hace patente la necesidad de incorporar más de 20 dominios representativos, para determinar el concepto semántico de cada segmento de texto (Kozareva et al. (2007)).

6.5 Integración de DRelevant en un sistema basado en aprendizaje

El sistema DRelevant ha sido utilizado para enriquecer el conjunto de características de un sistema de aprendizaje automático. El objetivo de este experimento es comprobar si la integración de la información proporcionada por el sistema DRelevant influye de forma positiva en el proceso de desambiguación. Para realizar el experimento, se han utilizado los datos de la tarea English lexical sample de SENSEVAL-2.

6.5.1 Sistema de aprendizaje inicial

El sistema de aprendizaje inicial para el desarrollo de nuestro experimento es el descrito en (Suárez (2004)). Este sistema de aprendizaje supervisado (WSD_MAX_ENT), está basado en el modelo probabilístico de máxima entropía (Ratnaparkhi (1998)), donde a partir del modelado de una serie de características o atributos y del aprendizaje a partir de corpus anotados semánticamente trata de resolver el problema de la ambigüedad semántica.

El conjunto de atributos utilizado por el sistema WSD_MAX_ENT se muestra en la Tabla 6.31.

Estos atributos se basan, principalmente, en el conocimiento lingüístico del contexto cercano a la palabra ambigua: palabras y composiciones de palabras que la acompañan, categorías gramaticales, rol gramatical, dependencias, etc.

La definición de atributos no tiene por qué ser exclusivamente automática y a partir del corpus de aprendizaje. También se podría incorporar información externa al corpus si fuera necesaria.

No relajados
0: la palabra ambigua
l: lemas (de palabras llenas) en ±1, ±2, ±3
s: palabras en posiciones ±1, ±2, ±3
b: lemas de pares de palabras en (-2, -1), (-1, +1), (+1, +2)
c: pares de palabras en (-2, -1), (-1, +1), (+1, +2)
p: categorías gramaticales de palabras en ±1, ±2, ±3
k_m: lemas de nombres que aparecen en al menos el m % de contextos de un sentido
r: rol gramatical de la palabra ambigua
d: la palabra de la que depende la ambigua
m: palabra compuesta a la que pertenece la ambigua
Relajados
L: lemas (de palabras llenas) en ±1, ±2, ±3
W: palabras llenas en ±1, ±2, ±3
S: palabras en ±1, ±2, ±3
B: lemas de pares de palabras en (-2, -1), (-1, +1), (+1, +2)
C: pares de palabras en (-2, -1), (-1, +1), (+1, +2)
P: categorías gramaticales en ±1, ±2, ±3
D: la palabra de la que depende la ambigua
M: palabra compuesta a la que pertenece la ambigua

Tabla 6.31. Conjunto de atributos de WSD_MAX_ENT

Y es en este punto donde añadimos la información proporcionada por el sistema DRelevant.

6.5.2 Nuevas características usando DRelevant

Para la incorporación de la información del sistema DRelevant en el sistema supervisado WSD_MAX_ENT se van a utilizar las etiquetas de WordNet Domains del contexto más cercano a la palabra a desambiguar. Concretamente se utilizarán las etiquetas de dominio de las dos palabras situadas a la derecha y a la izquierda de la palabra objetivo. Véase la Figura 6.11.

En este caso, la palabra objetivo es "car" y se utiliza el sistema DRelevant para anotar las cuatro palabras más cercanas a "car" con contenido semántico (nombres, verbos, adjetivos o adverbios) con sus respectivos dominios. Una vez acabado el proceso de desambiguación con DRelevant y anotadas las cuatro palabras

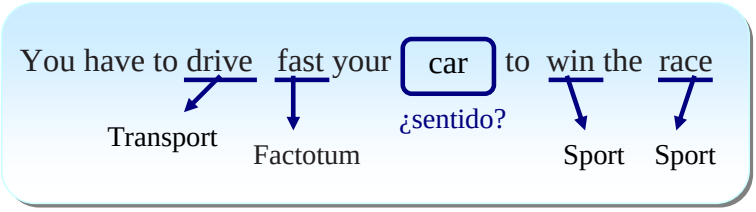


Figura 6.11. Anotación con DRelevant

con sus respectivos dominios, el segundo paso es integrar esta información en el corpus de entrenamiento del sistema, como una nueva característica más para el aprendizaje. El tercer paso es entrenar el sistema supervisado con esta nueva información. Para finalmente evaluar los resultados obtenidos con el corpus de test.

Cabe destacar que en el proceso de integración de los dominios como nuevos atributos del sistema, no se realizó ninguna codificación especial, sino que se utilizaron tal cual fueron anotados, lo que puede haber significado alguna merma en el acierto final.

6.5.3 Resultados

Los resultados tras el enriquecimiento con DRelevant se muestran en la Tabla 6.32. Los experimentos se realizaron sobre la tarea English Lexical Sample de Senseval-2 sobre una muestra de 29 nombres. En esta tabla se pueden comparar los resultados obtenidos antes y después de añadir los nuevos atributos.

Como se puede apreciar, de los 29 nombres escogidos sólo 4 empeoraron en el proceso de desambiguación tras usar los nuevos atributos. Los demás nombres o mantuvieron los mismos resultados o sufrieron una mejora tras el proceso de desambiguación.

La razón de que la mejora total sea de sólo un 2 % se debe principalmente a que gran parte de los nombres no incrementan su acierto. Además cabe destacar que el proceso de anotación del sistema DRelevant no es correcto al 100 %. Tampoco se se ha tenido en cuenta si el dominio "factotum" (etiqueta de WordNet Domains que indica que cierto nombre resulta inclasificable) debe eliminarse para mejorar el proceso de anotación. Incluso se podría probar una ventana de palabras mayor en torno al nombre

Nombres	SinDRelevant	ConDRelevant	Mejora
art	68,3	68,3	0
authority	53,8	56,3	2,5
bar	51,9	51	-0,9
bum	86,5	91,9	5,4
chair	89,8	89,8	0
channel	12,5	18,8	6,3
child	61	62,7	1,7
church	60	60	0
circuit	24,5	38,8	14,3
day	64	64,7	0,7
detention	90,9	86,4	-4,5
dyke	80	80	0
facility	71,4	64,3	-7,1
fatigue	86,8	86,8	0
feeling	60,4	66,7	6,3
grip	15,8	15,8	0
hearth	79,3	79,3	0
holiday	100	100	0
lady	87,5	90	2,5
material	36,2	51,7	15,5
mouth	56,9	58,8	1,9
nation	72	72	0
nature	43,2	46	2,8
post	51,2	51,2	0
restraint	48,4	51,6	3,2
sense	43,2	48,7	5,5
spade	82,4	88,2	5,8
stress	46	40,5	-5,5
yew	79,2	79,2	0
Total	62	64	2

Tabla 6.32. Enriquecimiento de un sistema basado en aprendizaje con DRelevant

objetivo. Además, también sería interesante aplicar estos atributos sobre verbos y adjetivos. Todas estas posibles mejoras quedan pendientes como trabajo futuro.

6.5.4 Test de McNemar

Tras la inclusión de los dominios como una nueva característica en el sistema supervisado WSD_MAX_ENT, es necesario determinar la significancia de los cambios producidos. Para ello, se va a utilizar el test de McNemar (Everitt (1977)) que determina si los resultados obtenidos antes y después de la inclusión de los dominios como nueva característica producen cambios significativos.

La Tabla 6.33 muestra la información utilizada para construir la tabla de contingencia:

Ejemplos clasificados erróneamente por ambos algoritmos	Ejemplos clasificados erróneamente por WSD_SinDOm pero no por WSD_ConDom
Ejemplos clasificados erróneamente por WSD_ConDom pero no por WSD_SinDOm	Ejemplos clasificados correctamente por ambos algoritmos

Tabla 6.33. Tabla de contingencia para el test de McNemar

Para abreviar, se utilizará la notación de la Tabla 6.34. Así se puede identificar cada caso mediante las variables: $n_{00}, n_{01}, n_{10}, n_{11}$.

n_{00}	n_{01}
n_{10}	n_{11}

Tabla 6.34. Notación abreviada Tabla de contingencia del test de McNemar

La Tabla 6.35 muestra los valores asociados a cada variable tras aplicar ambos algoritmos a las instancias de test de English Lexical Sample Senseval.

La suma de todas las variables es el número total de ejemplos en el conjunto de instancias de test. En nuestro caso: $447 + 96 + 68 + 685 = 1296$.

El test de McNemar se utiliza para comparar los resultados de una hipótesis nula o teórica H_0, con los resultados de la hipótesis para los valores reales observados H_1. La hipótesis nula tiene como

447	96
68	685

Tabla 6.35. Valores observados antes y después de la inclusión de los dominios

premisa que ambos algoritmos comparten los mismos errores, por tanto, $n_{01} = n_{10}$. Supongamos que el nivel de significancia es $\alpha = 0{,}05$ con 1 grado de libertad, por tanto, $\chi^2_{1-0{,}95} = 3{,}841459$.

Regla de decisión:

- Si $\chi^2 > \chi^2_{1-\alpha}$ rechazamos H_0 y existe asociación significativa (p-value $< \alpha$).
- Si $\chi^2 \leq \chi^2_{1-\alpha}$ asumimos H_0 y no existe asociación significativa (p-value $\geq \alpha$).

En nuestro caso en particular:

$$\chi^2 = \frac{(|n_{01}-n_{10}|-1)^2}{n_{01}+n_{10}} = \frac{(|96-68|-1)^2}{96+68} = 4{,}45$$

Como el valor de $\chi^2 = 4{,}45 > \chi^2_{1-0{,}95} = 3{,}84$ entones $p < 0{,}05$, concluyendo que la utilización de dominios es significativa para la mejora del sistema WSD_MAX_ENT.

Universitat d'Alacant
Universidad de Alicante

Capítulo 7

Conclusiones y trabajos futuros

En este capítulo final y como consecución de esta Tesis doctoral, se presenta una síntesis sobre el trabajo desarrollado, un análisis de los beneficios aportados por la desambiguación automática a otras de tareas de PLN, una serie de propuestas con vistas al futuro, así como el conjunto de publicaciones relevantes derivadas de este trabajo.

7.1 Aportaciones

En esta Tesis se ha presentado la definición y evaluación de varios métodos de resolución de la ambigüedad semántica: DRelevant, DLSA y SenseDiscrim. Todos los métodos presentados se clasifican dentro de la categoría de métodos no supervisados, basados en conocimiento. Estos métodos han sido evaluados según las especificaciones de la competición Senseval mostrando una comparativa respecto a otros sistemas. Además, se han integrado una serie de recursos semánticos (Dominios Relevantes, SUMO) sobre diferentes técnicas (LSA, Machine Learning) con el objetivo de resolver problemas que afectan a otras tareas de PLN, tales como: reconocimiento de la variabilidad semántica o detección y clasificación de nombres propios. Además, como consecución del estudio de la distribución y relaciones entre los sentidos en bases de datos léxicas como WordNet se ha creado un nuevo recurso

léxico: Dominios Relevantes. Este recurso es susceptible de integrarse en otros sistemas de WSD o servir de referencia semántica para otras tareas de PLN.

Las principales aportaciones siguiendo la estructura de esta Tesis han sido:

7.1.1 Estudio del estado del arte

Estudio de la evolución de los sistemas de resolución automática de la ambigüedad desde los comienzos del PLN hasta la fecha actual. Se ha realizado una clasificación de dichos sistemas dentro de distintas categorías, distinguiendo entre sistemas supervisados y sistemas no supervisados.

A la vista de los resultados obtenidos en las distintas competiciones para evaluación de sistemas que resuelven de forma automática la ambigüedad en el lenguaje, los sistemas supervisados han demostrado ser más eficientes que los sistemas no supervisados. Sin embargo, el principal problema de los sistemas supervisados reside en la escasez de corpus anotados para su entrenamiento.

7.1.2 Estudio de los sistemas de evaluación en WSD

Se ha realizado un estudio de la evolución de los métodos de evaluación de sistemas de WSD. Originalmente existía una imposibilidad de realizar estudios comparativos entre diferentes sistemas debido a la utilización de distintos tipos de anotación semántica o de corpus utilizados para la evaluación. Ante este problema, desde 1998 hasta la actualidad, la evaluación de sistemas de WSD se realiza bajo un marco común de evaluación: la competición Senseval. Esta competición unifica criterios de evaluación permitiendo de esta forma evaluar distintos sistemas y comparar los resultados obtenidos.

7.1.3 Descripción de los recursos léxicos utilizados

Se han descrito los distintos recursos léxicos utilizados para el desarrollo de los métodos de WSD presentados en este trabajo.

En nuestro caso, la base de conocimiento principal sobre la que subyacen todos los métodos desarrollados es WordNet: una base de datos léxica que sigue una serie de criterios psicolingüísticos, ampliamente utilizada en PLN. A partir de esta base de datos léxica se han utilizado otras ontologías como SUMO o WordNet Domains que enriquecen las inter-relaciones de palabras presentes en WordNet. Además, mediante la utilización de una versión extendida de WordNet, Extended WordNet, se ha mejorado la obtención del nuevo recurso léxico Dominios Relevantes. Este recurso léxico proporciona información relevante acerca de las palabras polisémicas junto con las categorías semánticas con las que se relacionan.

Además, se ha presentado y descrito detalladamente la técnica de LSA, que permite extraer relaciones existentes entre palabras a través de sus ocurrencias en diferentes contextos. Esta técnica ha sido adaptada a nuestras necesidades, transformando el concepto de contextos (documentos) en categorías semánticas (dominios).

7.1.4 Definición de los métodos evaluados

Se han descrito los distintos métodos de WSD, los cuales, utilizan la información de WordNet, SUMO y la técnica de LSA. Estos métodos pueden aplicarse a diferentes lenguas utilizando el enlace ILI de EuroWordNet. Únicamente se requiere un preproceso inicial de los textos para obtener los lemas de las palabras del contexto ambiguo. Asimismo, mediante la utilización de la información de Extended WordNet se han mejorado los resultados obtenidos en el proceso de desambiguación.

Todos los métodos definidos han sido evaluados sobre los corpus de Senseval y comparados con el resto de sistemas participantes en las diferentes tareas.

7.1.5 Evaluación y aplicación de los sistemas de WSD a tareas de PLN

Se ha presentado el marco de evaluación de sistemas de WSD Senseval, en sus diferentes ediciones. En cada edición se han

mantenido una serie de tareas relacionadas con WSD (*All Words* y *Lexical Sample* en diferentes idiomas) y además se han ido añadiendo progresivamente otro tipo de tareas en las que el proceso de desambiguación automática es beneficioso (Web People Search, desambiguación de preposiciones, detección de sentimientos, etc).

Todos los métodos definidos han sido evaluados siguiendo los criterios de SENSEVAL. Además, se ha evaluado la integración de estos sistemas no supervisados con otros sistemas de WSD, obteniendo buenos resultados.

Dado que la tarea de WSD, no está considerada como una tarea final, así como, traducción automática o clasificación de documentos, se han realizado una serie de experimentos aplicando los sistemas de WSD sobre otras tareas de PLN. En nuestro caso, se han aplicado para resolver la implicación textual, la detección de paráfrasis o la clasificación de nombres propios comunes pertenecientes a distintas personas.

7.2 Trabajos Futuros

Como trabajos futuros queda pendiente la elaboración de un sistema de WSD que combine los recursos obtenidos en esta Tesis para crear un sistema de WSD supervisado. El objetivo es determinar si el modelado de características utilizando un sistema de WSD no supervisado basado en conocimiento, ayuda a un sistema de WSD supervisado. En (Vázquez et al. (2007)) se describe la propuesta de este sistema.

También queda pendiente mejorar el recurso Relevant Domains incluyendo información relativa a las relaciones existentes entre los dominios y la jerarquía de relaciones de WordNet: hiponimina, meronimia, hiperonimia... Así como mejorar la obtención de la matriz conceptual utilizada por LSA, incorporando información relativa a las relaciones existentes entre palabras: verbos con objetos directos, sintagmas nominales, etc. Además, el sistema SenseDiscrim puede ser enriquecido mediante la obtención de nue-

vos patrones que incorporen información relacionada con verbos y adjetivos.

Por último, se está desarrollando la adaptación de los métodos descritos en esta Tesis para su aplicación a la resolución de los tests de TOEFL que tienen como objetivo la detección de similitudes entre pares de palabras. El objetivo de este estudio es determinar si ante la falta de contexto proporcionado, se pueden adaptar los recursos existentes para seleccionar pares de palabras fuertemente relacionadas semánticamente.

7.3 Producción científica

A continuación se muestran las publicaciones realizadas como consecución de esta Tesis. Todas ellas en orden cronológico desde 2002 hasta 2008.

2008

Zornitsa Kozareva, Sonia Vázquez, Andrés Montoyo. **Domain Information for Fine-Grained Person Name Categorization.** CICLing 2008. Haifa (Israel). pp: 311-321. *Lecture Notes in Computer Science.* Vol: 4919/2008. ISSN: 0302-9743.

2007

Zornitsa Kozareva, Sonia Vázquez, Andrés Montoyo. **The influence of context during the categorization and discrimination of Spanish and Portuguese person names.** SEPLN 2007. Sevilla (España). pp: 81-88. *Procesamiento del Lenguaje Natural.* Vol: 39. ISSN: 1135-5948.

Zornitsa Kozareva, Sonia Vázquez, Andrés Montoyo. **The Usefulness of Conceptual Representation for the Identification of Semantic Variability Expressions.** CICLing 2007. Mexico city (Mexico). pp: 325-336. *Lecture Notes in Computer Science.* Vol: 4394/2007. ISSN: 0302-9743.

Universitat d'Alacant
Universidad de Alicante

Sonia Vázquez, Andrés Montoyo, Zornitsa Kozareva. **Word Sense Disambiguation Using Extended Relevant Domains Resource.** IC-AI 2007. Las Vegas (Nevada, USA). pp: 823-828. *CSREA Press.* ISBN: 1-60132-024-8.

Sonia Vázquez, Zornitsa Kozareva, Andrés Montoyo. **How Context and Semantic Information Can Help a Machine Learning System?** MICAI 2007. Aguascalientes (Mexico). pp: 996-1003. *Lecture Notes in Computer Science.* Vol: 4827/2007 ISSN: 0302-9743.

Zornitsa Kozareva, Sonia Vázquez, Andrés Montoyo. **Multilingual Name Disambiguation with Semantic Information.** TSD 2007. Pilsen (República Checa). pp: 23-30. *Lecture Notes in Computer Science.* Vol: 4629/2007. ISSN: 0302-9743.

Zornitsa Kozareva, Sonia Vázquez, Andrés Montoyo. **Discovering the Underlying Meanings and Categories of a Name through Semantic and Domain Information.** Recent Advances in Natural Language Processing (RANLP 2007). Borovets (Bulgaria).

Zornitsa Kozareva, Sonia Vázquez, Andrés Montoyo. **A Language Independent Approach for Name Categorization and Discrimination.** ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. 45th Annual Meeting of the Association for Computational Linguistics. Praga (República Checa).

Zornitsa Kozareva, Borja Navarro, Sonia Vázquez, Andrés Montoyo. **UA-ZBSA: A Headline Emotion Classification through Web Information.** International Workshop on Semantic Evaluations SEMEVAL. 4th Interntional SemEval-ACL 2007. Praga (República Checa).

Zornitsa Kozareva, Sonia Vázquez, Andrés Montoyo. **UA-ZSA: Web Page Clustering on the basis of Name Di-**

sambiguation. International Workshop on Semantic Evaluations SEMEVAL. 4th Interntional SemEval- ACL 2007. Praga (República Checa).

2006

Sonia Vázquez, Zornitsa Kozareva, Andrés Montoyo. **Contribución de la información semántica en un sistema de aprendizaje automático para resolver la implicación textual.** SEPLN 2006. Zaragoza (España). pp: 189-196. *Procesamiento del Lenguaje Natural.* Vol: 37. ISSN: 1135-5948.

Zornitsa Kozareva, Sonia Vázquez, Andrés Montoyo. **University of Alicante at QA@CLEF2006: Answer Validation Exercise.** CLEF 2006. Alicante (España). pp: 522-525. *Lecture Notes in Computer Science.* Vol: 4730/2007. ISSN: 0302-9743.

Sonia Vázquez, Zornitsa Kozareva, Andrés Montoyo. **Textual Entailment Beyond Semantic Similarity Information.** MICAI 2006. Apizaco (Mexico). pp: 900-910. *Lecture Notes in Computer Science.* Vol: 4293/2006. ISSN: 0302-9743.

Zornitsa Kozareva, Sonia Vázquez, Andrés Montoyo. **The Effect of Semantic Knowledge Expansion to Textual Entailment Recognition.** TSD 2006. Brno (República Checa). pp: 143-150. *Lecture Notes in Computer Science.* Vol: 4188/2006. ISSN: 0302-9743.

2005

Borja Navarro, Lorenza Moreno-Monteagudo, Elisa Noguera, Sonia Vázquez, Fernando Llopis, Andrés Montoyo. **"How Much Context Do You Need?": An Experiment About the Context Size in Interactive Cross-Language Question Answering.** CLEF 2005. Viena (Austria). pp: 273-282. *Lecture Notes in Computer Science.* Vol: 4022/2006. ISSN:

0302-9743.

2004

Iulia Nica, Maria Antònia Martí, Andrés Montoyo, Sonia Vázquez. **Combining EWN and Sense-Untagged Corpus for WSD.** CICLing 2004. Seúl (Korea). pp: 188-200. *Lecture Notes in Computer Science.* Vol: 2945/2004. ISSN: 0302-9743.

Sonia Vázquez, Estela Saquete, Andrés Montoyo, Patricio Martínez-Barco, Rafael Muñoz. **The Role of Temporal Expressions in Word Sense Disambiguation.** CICLing 2004. Seúl (Korea). pp: 209-212. *Lecture Notes in Computer Science.* Vol: 2945/2004. ISSN: 0302-9743.

Borja Navarro, Lorenza Moreno, Sonia Vázquez, Fernando Llopis, Andrés Montoyo, Miguel Angel Varó. **Improving Interaction with the User in Cross-Language Question Answering Through Relevant Domains and Syntactic Semantic Patterns.** CLEF 2004. Bath (Reino Unido). pp: 334-342. *Lecture Notes in Computer Science.* Vol: 3491/2005. ISSN: 0302-9743.

Sonia Vázquez, Andrés Montoyo, German Rigau. **Using Relevant Domains Resource for Word Sense Disambiguation**. IC-AI 2004. Las Vegas (Nevada, USA). pp: 784-789. *CSREA Press.* ISBN: 1-932415-32-7.

Iulia Nica, Andrés Montoyo, Sonia Vázquez, Maria Antònia Martí. **An Unsupervised WSD Algorithm for a NLP System.** NLDB 2004. Manchester (Reino Unido). pp: 288-298. *Lecture Notes in Computer Science.* Vol: 3136/2004. ISSN: 0302-9743.

Iulia Nica, Maria Antonia Martí, Andrés Montoyo, Sonia Vázquez. **Intensive Use of Lexicon and Corpus for WSD.** SEPLN 2004. Barcelona (España). pp: 147-154. *Pro-*

cesamiento del Lenguaje Natural. Vol: 32. ISSN: 1135-5948.

Sonia Vázquez, Andrés Montoyo. **Multilingual Extended WordNet.** Biennial Iberoamerican Conference on Artificial Intelligence (IBERAMIA). Puebla (Mexico).

Sonia Vázquez, Andrés Montoyo. **Utilización del recurso Dominios Relevantes en WSD.** III Jornadas en Tecnología del Habla. Valencia (España).

Sonia Vázquez, Rafael Romero, Armando Suárez, Andrés Montoyo, Manuel García, Maria Teresa Martín, Miguel Ángel García, Alfonso Ureña, Davide Buscaldi, Paolo Rosso, Antonio Molina, Ferran Plá, Encarna Segarra. **The R2D2 Team at SENSEVAL-3.** International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL) in conjunction with the Annual Meeting of Association for Computational Linguistics. Barcelona (España).

Sonia Vázquez, Rafael Romero, Armando Suárez, Andrés Montoyo, Iulia Nica, Maria Antonia Martí. **The University of Alicante systems at SENSEVAL-3.** International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL) in conjunction with the Annual Meeting of Association for Computational Linguistics. Barcelona (España).

Iulia Nica, Maria Antonia Martí, Andrés Montoyo, Sonia Vázquez. **Enriching EWN with syntagmatic information by means of WSD.** International Conference on Language Resources and Evaluation (LREC). Lisboa (Portugal).

Iulia Nica, Maria Antonia Martí, Andrés Montoyo, Sonia Vázquez. **Towards filling the gap between lexicon and corpus.** International Conference on Language Resources and Evaluation (LREC). Lisboa (Portugal).

Universidad de Alicante

2003

Sonia Vázquez, Andrés Montoyo, German Rigau. **Método de desambiguación léxica basada en el recurso léxico Dominios Relevantes.** SEPLN 2003. Alcalá de Henares (España). pp: 141-148. *Procesamiento del Lenguaje Natural.* Vol: 31. ISSN: 1135-5948.

2002

Sonia Vázquez, Ma Carmen Calle, Susana Soler, Andrés Montoyo. **Specification Marks Method: Design and Implementation.** CICLing 2002. Mexico city (Mexico). pp: 439-442. *Lecture Notes in Computer Science.* Vol: 2276/2002. ISSN: 0302-9743.

Andrés Montoyo, Rafael Romero, Sonia Vázquez, Mª Carmen Calle, Susana Soler. **The Role of WSD for Multilingual Natural Language Applications.** TSD 2002. Brno (República Checa). pp: 41-48. *Lecture Notes in Computer Science.* Vol: 2448/2002. ISSN: 0302-9743 .

Universitat d'Alacant
Universidad de Alicante

Apéndice A

Acrónimos

ACL. Association of Computational Linguistics.
ACM. Association for Computing Machinery.
AR. Association Ratio.
AVE. Answer Validation Exercise.
BNC. British National Corpus.
CBC. Clustering by Committee.
CICLING. Conference on Intelligent Text Processing and Computational Linguistics.
CLEF. Cross Language Evaluation Forum.
COLING. International Conference on Computational Linguistics.
DARPA. Defense Advanced Research Project Agency.
EI. Extracción de Información.
EWN. EuroWordNet.
GATE. General Architecture for Text Engineering.
GPLSI. Grupo de Procesamiento del Lenguaje Natural.
ICAI. International Conference on Artificial Intelligence.
iCLEF. Interactive track of CLEF
ILI. Inter Lingual Index
IM. Información Mutua.
KIF. Knowledge Interchange Format.
LCS. Lowest Common Subsumer.
LongCS. Longest Common Subsequence.

Universidad de Alicante

LDOCE. Longman Dictionary of Contemporary English.
LREC. International Conference on Language Resources and Evaluation.
LSA. Latent Semantic Analysis o Análisis de la Semántica Latente.
LSI. Latent Semantic Indexing.
LVQ. Learning Vector Quantification.
MI. Mutual Information.
MICAI. Mexican International Conference on Artificial Intelligence.
MUC. Message Understanding Conferences.
MRD. Machine Readable Dictionary.
NBC. Naïve Bayes Classifier.
NLDB. International Conference on Applications of Natural Language to Information Systems.
PASCAL. Pattern Analysis, Statistical Modelling and Computational Learning.
PLN. Procesamiento del Lenguaje Natural.
PMI. Pointwise Mutual Information.
QA. Question Answering.
RA. Ratio de asociación.
RAE. Real Academia Española.
RANLP. Recent Advances in Natural Language Processing.
RI. Recuperación de Información.
RNA. Red Neuronal Artificial.
RTE. Recognising Textual Entailment.
SemCor (SEMantic COncoRdance)
Senseval/Semeval. Evaluation Exercises for the Semantic Analysis of Text.
SEPLN. Sociedad Española para el Procesamiento del Lenguaje Natural.
SFC. Subject Field Codes.
SIGIR. Special Interest Group on Information Retrieval.
SIGLEX. Special Interest Group on the Lexicon of the Association for Computational Linguistics.
SUMO. Suggested Upper Merged Ontology.
SVD. Singular Value Decomposition.

Universitat d'Alacant
Universidad de Alicante

SVM. Support Vector Machines.
TA. Traducción automática.
TOEFL. Test of English as a Foreign Language.
TREC. Text Retrieval Conference.
WDD. Word Domain Disambiguation.
WePS. Web People Search.
WND. WordNet Domains.
WSD. Word Sense Disambiguation.
WSD_MAX_ENT. Sistema de desambiguación automática basado en modelos de probabilidad de máxima entropía.

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Bibliografía

Abney, Steven P. (2002). «Bootstrapping», en *ACL*, págs. 360–367.

Abney, Steven P. (2004). «Understanding the yarowsky algorithm», *Computational Linguistics*, **30**(3), 365–395.

ACL, editor (2001). *Proceedings of the SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*.

ACL, editor (2004). *Proceedings of the SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*.

ACL, editor (2007). *Proceedings of the SEMEVAL-2007: Fourth Internatinal Workshop on Semantic Evaluations*.

Agirre, E. y D. Martínez (2000). «Exploring automatic word sense disambiguation with decision lists and the Web», en *Proceedings of the Semantic Annotation And Intelligent Annotation workshop organized by COLING*, Luxembourg.

Agirre, Eneko, Itziar Aldabe, Mikel Lersundi, David Martínez, Eli Pociello y Larraitz Uria (2004). «The basque lexical-sample task», en Rada Mihalcea y Phil Edmonds, editores, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, págs. 1–4, Association for Computational Linguistics, Barcelona, Spain.

Agirre, Eneko, Bernardo Magnini, Oier Lopez de Lacalle, Arantxa Otegi, German Rigau y Piek Vossen (2007). «Semeval-2007 task 01: Evaluating wsd on cross-language information retrieval», en *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, págs. 1–6, Association for Computational Linguistics, Prague, Czech Republic.

Agirre, Eneko y David Martinez (2001). «Learning class-to-class selectional preferences», en *Proceedings of the Workshop Çomputational Natural Language Learning"(CoNLL-2001). In conjunction with ACL'2001/EACL'2001*, Toulouse, France.

Agirre, Eneko y German Rigau (1996). «Word Sense Disambiguation using Conceptual Density», en *Proceedings of the 16th International Conference on Computational Linguistic (COLING´96)*, Copenhagen, Denmark.

Agirre, Eneko y Aitor Soroa (2007). «Semeval-2007 task 02: Evaluating word sense induction and discrimination systems», en *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, págs. 7–12, Association for Computational Linguistics, Prague, Czech Republic.

Akhmatova, E. (2005). «Textual entailment resolution via atomic propositions», en *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment, 2005.*, págs. 61–64.

Allen, James F. (1984). «Towards a general theory of action and time», *Artif. Intell.*, **23**(2), 123–154.

Andreevska, A., Z. Li y S. Bergler (2005). «Can shallow predicate argument structure determine entailment?», en *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment, 2005.*, págs. 45–48.

Artiles, Javier, Julio Gonzalo y Satoshi Sekine (2007). «The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task», en *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, págs. 64–69, Association for Computational Linguistics, Prague, Czech Republic.

Atkins, Sue (1992). «Tools for computer-aided corpus lexicography: the hector project», págs. 1–60.

Atserias, J., J. Carmona, I. Castellón, S. Cervell, M. Civit, L. Màrquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé y J. Turmo (1998). «Morphosyntactic Analysis and Parsing of Unrestricted Spanish Text», en *Proceedings of First International Conference on Language Resources and Evaluation (LREC'98)*, Granada, Spain.

Baker, Collin, Michael Ellsworth y Katrin Erk (2007). «Semeval-2007 task 19: Frame semantic structure extraction», en *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, págs. 99–104, Association for Computational Linguistics, Prague, Czech Republic.

Banerjee, Satanjeev y Ted Pedersen (2002). «An adapted lesk algorithm for word sense disambiguation using wordnet.», en *CICLing*, págs. 136–145.

Barzilay, Regina y Kathleen McKeown (2001). «Extracting paraphrases from a parallel corpus», en *ACL*, págs. 50–57.

Barzilay, Regina y Kathleen McKeown (2003). «Learning to paraphrase: An unsupervised approach using multiple-sequence alignment», en *HTLT-NAACL*, págs. 16–23.

Bloehdorn, Stephan y Andreas Hotho (2004). «Text classification by boosting weak learners based on terms and concepts.», en *ICDM*, págs. 331–334.

Blum, Avrim y Tom M. Mitchell (1998). «Combining labeled and unlabeled sata with co-training», en *COLT*, págs. 92–100.

Boguraev, B. y T. Briscoe (1989). *Computational lexicography for Natural Language Processing*, Longman, London and New York.

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

Brockmann, Carsten y Mirella Lapata (2003). «Evaluating and combining approaches to selectional preference acquisition», en *EACL*, págs. 27–34.

Brown, Peter F., Stephen A. Della Pietra y Vincent J. Della Pietra (1991). «Word sense Disambiguation using statistical methods», en *Proceedings of 29th Annual Meeting of the Association for Computational Linguistics*, págs. 264–270.

Calzolari, Nicoletta, Ornella Corazzari y Antonio Zampolli (2001). «Lexical-semantic tagging of an italian corpus», en *CICLing*, págs. 291–304.

Cederberg, Scott y Dominic Widdows (2003). «Using lsa and noun coordination information to improve the precision and recall of automatic hyponymy extraction», en *In Proceedings of CoNLL*, págs. 111–118.

Chklovski, Timothy y Rada Mihalcea (2002). «Building a sense tagged corpus with open mind word expert», en *Proceedings of the ACL-02 workshop on Word sense disambiguation*, págs. 116–122, Association for Computational Linguistics, Morristown, NJ, USA.

Chklovski, Timothy, Rada Mihalcea, Ted Pedersen y Amruta Purandare (2004). «The senseval-3 multilingual englishhindi lexical sample task», en Rada Mihalcea y Phil Edmonds, editores, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, págs. 5–8, Association for Computational Linguistics, Barcelona, Spain.

Church, Kenneth, William Gale, Patrick Hanks y Donald Hindle (1991). «Using statistics in lexical analysis», en Uri Zernik, editor, *Lexical acquisition*, págs. 115–164, Erlbaum.

Church, Kenneth Ward y Patrick Hanks (1990). «Word association norms, mutual information, and lexicography», *Computational Linguistics*, **16**(1), 22–29.

Ciaramita, Massimiliano y Mark Johnson (2000). «Explaining away ambiguity: Learning verb selectional preference with bayesian networks», en *COLING*, págs. 187–193.

Civit, Montse (2003). «Criterios de etiquetación y desambiguación morfosintáctica de corpus en español», Tesis Doctoral. Universidad de Barcelona.

Clough, Paul y Mark Stevenson (2004). «Cross-language information retrieval using eurowordnet and word sense disam-

biguation.», en *ECIR*, págs. 327–337.
Cohen, Jacob (1960). «A coefficient of agreement for nominal scales», *Educational and Psychological Measurement*, **20**(1), 37–46.
Connine, Cynthia (1990). «Effects of sentence context and lexical knowledge in speech processing», págs. 281–294.
Corazzari, Ornella y Antonietta Alonge (2001). «Italwordnet: extending and exploiting an existing resource for computational tasks», .
Corley, Courtney y Rada Mihalcea (2005). «Measures of text semantic similarity», en *ACL workshop on Empirical Modeling of Semantic Equivalence*.
Cover, Thomas M. y Joy A. Thomas (1991). *Elements of information theory*, Wiley-Interscience, New York, NY, USA.
Cowie, Jim, Joe Guthrie y Louise Guthrie (1992). «Lexical disambiguation using simulated annealing», en *Proceedings of the 14th International Conference on Computational Linguistic, COLING´92*, págs. 359–365, Nantes, France.
Cuadros, Montse y German Rigau (2007). «Semeval-2007 task 16: Evaluation of wide coverage knowledge resources», en *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, págs. 81–86, Association for Computational Linguistics, Prague, Czech Republic.
Cunningham, H. (2002). «GATE, a General Architecture for Text Engineering», *Computers and the Humanities*, **36**, 223–254.
Cunningham, H. (2005). «Information Extraction, Automatic», *Encyclopedia of Language and Linguistics, 2nd Edition*.
D., Harman (1995). «Overview of the 3rd text retrieval conference (trec-3)», National Institute of Standards and Technology (NIST) Special Publication 500-225, US.
D., Harman (1996). «Overview of the 3rd text retrieval conference (trec-4)», National Institute of Standards and Technology (NIST) Special Publication 500-236, US.
Dagan, I., A. Itai y U. Schwall (1991). «Two languages are more informative than one», *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, págs. 130–137.

Dagan, Ido, Oren Glickman y Bernardo Magnini (2005). «The pascal recognising textual entailment challenge.», en *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*.

Dagan, Ido, Lillian Lee y Fernando C.Ñ. Pereira (1997). «Similarity-based methods for word sense disambiguation», *CoRR*, **cmp-lg/9708010**.

Dagan, Ido, Lillian Lee y Fernando C.Ñ. Pereira (1999). «Similarity-based models of word cooccurrence probabilities», *Machine Learning*, **34**(1-3), 43–69.

Dagan, Ido, Shaul Marcus y Shaul Markovitch (1993). «Contextual word similarity and estimation from sparse data», en *In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, págs. 164–171.

Dagan, Ido, Fernando C.Ñ. Pereira y Lillian Lee (1994). «Similarity-based estimation of word cooccurrence probabilities», en *ACL*, págs. 272–278.

Deerwester, Scott C., Susan T. Dumais, Thomas K. Landauer, George W. Furnas y Richard A. Harshman (1990). «Indexing by latent semantic analysis», *JASIS*, **41**(6), 391–407.

Diab, Mona, Musa Alkhalifa, Sabry ElKateb, Christiane Fellbaum, Aous Mansouri y Martha Palmer (2007). «Semeval-2007 task 18: Arabic semantic labeling», en *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, págs. 93–98, Association for Computational Linguistics, Prague, Czech Republic.

Dill, Stephen, Nadav Eiron, David Gibson, Daniel Gruhl, Ramanathan V. Guha, Anant Jhingran, Tapas Kanungo, Sridhar Rajagopalan, Andrew Tomkins, John A. Tomlin y Jason Y. Zien (2003). «Semtag and seeker: bootstrapping the semantic web via automated semantic annotation.», en *WWW*, págs. 178–186.

Dolan, Bill, Chris Quirk y Chris Brockett (2004). «Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources», en *Proceedings of the 20th*

International Conference on Computational Linguistics, Geneva, Switzerland.

Duda, Richard, Peter Hart y David Stork (2001). «Pattern classification», Segunda Edición. Nueva York: John Wiley & Sons.

Edmonds, Philip y Adam Kilgarriff (1998). «Introduction to the special issue on evaluating word sense disambiguation systems.», en *Journal of Natural Language Engineering*.

Escudero, G., L. Márquez y G. Rigau (2000). «Naive Bayes and Exemplar-Based approaches to Word Sense Disambiguation Revisited», en *Proceedings of the 14th European Conference on Artificial Intelligence, ECAI-2000*, Berlin, Germany.

Everitt, B. S. (1977). *The analysis of contingency tables*, Chapman and Hall.

Fellbaum, Christiane (1998). *WordNet: An Electronic Lexical Database*, The MIT Press.

Firth, J. R. (1957). «A synopsis of linguistic theory. studies in linguistic analysis», en *Special Volume, Philological Society*, págs. 1–32.

Frakes, William B. y Ricardo A. Baeza-Yates, editores (1992). *Information Retrieval: Data Structures & Algorithms*, Prentice-Hall.

Framis, Francesc Ribas (1994). «An experiment on learning appropriate selectional restrictions from a parsed corpus», en *COLING*, págs. 769–774.

Francis, W.Ñ. y H. Kucera (1979). «Brown corpus manual», *inf. téc.*, Department of Linguistics, Brown University, Providence, Rhode Island, US.

Furnas, George W., Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, Richard A. Harshman, Lynn A. Streeter y Karen E. Lochbaum (1988). «Information retrieval using a singular value decomposition model of latent semantic structure», en *SIGIR*, págs. 465–480.

G. Miller, T. Randee, C. Leacock y R. Bunker (1993). «A Semantic Concordance», en *Proceeding of 3rd DARPA Workshop on Human Language Tecnology*, págs. 303–308, Plainsboro, New Jersey.

Gale, W., K. Church y D. Yarowsky (1992a). «Estimating upper and lower bounds on the performance or word-sense disambiguation programs», en *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistic*.

Gale, William, Kenneth Church y David Yarowsky (1992b). «A method for disambiguating word senses in a large corpus», *Computers and the Humanities*, **26**, 415–439.

Gan, Kok Wee y Ping Wai Wong (2000). «Annotating information structures in chinese texts using hownet», en *Proceedings of the second workshop on Chinese language processing*, págs. 85–92, Association for Computational Linguistics, Morristown, NJ, USA.

García, Manuel (2006). «Resolución de la ambigüedad léxica mediante aprendizaje por cuantificación vectorial», Tesis Doctoral. Departamento de Informática. Universidad de Jaén.

Genesereth, Michael R. (1991). «Knowledge interchange format», en *KR*, págs. 599–600.

Gildea, Daniel y Daniel Jurafsky (2002). «Automatic labeling of semantic roles», *Computational Linguistics*, **28**(3), 245–288.

Girju, Roxana, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney y Deniz Yuret (2007). «Semeval-2007 task 04: Classification of semantic relations between nominals», en *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, págs. 13–18, Association for Computational Linguistics, Prague, Czech Republic.

Golding, A. R. (1995). «A bayesian-hybrid method for context-sensitive spelling correction», In Proceedings of the 3rd Workshop on Very Large Corpora, ACL.

Gonzalo, Julio y Douglas W. Oard (2004). «iclef 2004 track overview: Pilot experiments in interactive cross-language question answering», en *CLEF*, págs. 310–322.

Graesser, Arthur, Danielle McNamara, Max Louwerse y Zhiqiang Cai (2004). «Coh-metrix: Analysis of text on cohesion and language.», **36**(2), 193–202.

Group, CORPORATE PDP Research (1986). *Parallel distributed processing: explorations in the microstructure of cogni-*

tion, vol. 2: psychological and biological models, MIT Press, Cambridge, MA, USA.

Grozea, Cristian (2004). «Finding optimal parameter settings for high performance word sense disambiguation», en Rada Mihalcea y Phil Edmonds, editores, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, págs. 125–128, Association for Computational Linguistics, Barcelona, Spain.

Haley, Debra Trusso, Pete Thomas, Anne De Roeck y Marian Petre (2005). «A research taxonomy for latent semantic analysis-based educational applications», en *In Proceedings of the International Conference on Recent Advances in Natural Language Processing*, págs. 21–23.

Halliday, Michael y Ruqaiya Hasan (1976). «Cohesion in English», *London:Longman*.

Harabagiu, Sanda M., George A. Miller y Dan I. Moldovan (1999). «Wordnet 2 - a morphologically and semantically enhanced resource», en *SIGLEX*.

Harris, Zellig (1968). «Mathematical structures of language», en *New York: Interscience Publishers John Wiley & Sons*.

Hawkins, Paul y David Nettleton (2000). «Large scale wsd using learning applied to senseval», en *Computers and the Humanities*, vol. 34, págs. 135–140.

Herrera, J., A. Peñas y F. Verdejo. (2005). «Textual entailment recognition based on dependency analysis and wordnet.», en *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment,2005.*

Hirst, Graeme y David St-Onge (1998). «Lexical chains as representations of context in the detection and correction of malaproprisms», *In WordNet: An electronic lexical database, ed. by Christiane Fellbaum*, págs. 305–332.

HLT1 (1993). «Proceedings of the arpa workshop on human language technology», Morgan Kaufmann, San Mateo, CA.

Jaccard, P. (1901). «Étude comparative de la distribution florale dans une portion des alpes et des jura», *Bull Soc Vaudoise Sci Nat*, **37**, 547–579.

Jiang, Jay J. y David W. Conrath (1997). «Semantic similarity based on corpus statistics and lexical taxonomy», *CoRR*, **cmp-lg/9709008**.

Jin, Peng, Yunfang Wu y Shiwen Yu (2007). «Semeval-2007 task 05: Multilingual chinese-english lexical sample», en *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, págs. 19–23, Association for Computational Linguistics, Prague, Czech Republic.

Jing, Hongyan y Evelyne Tzoukermann (1999). «Information retrieval based on context distance and morphology», en *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, págs. 90–96, ACM Press, New York, NY, USA.

Kilgariff, A. (2001). «English lexical sample task description», *Proceedings of Senseval-2: Second International Workshop on Evaluating Word Sense Disambiguaion Systems.*, págs. 17–20.

Kilgarriff, A. (1998a). «Gold standard for evaluating word sense disambiguation programs», *Computer Speech and Language, Special Issue on evaluation*, **12**(3).

Kilgarriff, A. (1998b). «Senseval: An exercise in evaluating word sense disambiguation programs», en *LREC, Granada, May 1998*, págs. 581–588.

Kilgarriff, A. y M. Palmer (2000). «Introduction to the Special Issue on SENSEVAL», *Computers and the Humanities*, **34(1/2)**(1-13).

Kilgarriff, A. y J Rosenzweig (2000). «Framework and results for english Senseval», *Computers and the Humanities*, **34**(1-2).

Kohonen, T. (1989). *Self-organization and associative memory: 3rd edition*, Springer-Verlag New York, Inc., New York, NY, USA.

Kozareva, Zornitsa y Andrés Montoyo (2006). «Mlent: The machine learning entailment system of the university of alicante», en *In Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment.*

Universitat d'Alacant
Universidad de Alicante

Kozareva, Zornitsa, Sonia Vázquez y Andrés Montoyo (2006). «Adaptation of a machine learning textual entailment system to a multilingual answer validation exercise», en *In Working Notes of CLEF 2006*.

Kozareva, Zornitsa, Sonia Vázquez y Andrés Montoyo (2007). «The usefulness of conceptual representation for the identification of semantic variability expressions», en *CICLing*, págs. 325–336.

Krovets, R. (1997). «Homonymy and polysemy in information retrival», en *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistic and 8th Conference of the European Chapter of the Association for Computational Linguistic*, págs. 72–79, Madrid, Spain.

Krovets, R. y W. Bruce Croft (1992). «Lexical ambiguity and information retrieval», *ACM*, **10**(2), 115–141.

Krovetz, Robert (1998). «More than one sense per discourse», en *Proceedings of SENSEVAL Workshop*.

Landauer, Thomas K. y Susan T. Dumais (1997). «A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge», en *Psychological Review*, págs. 211–240.

Landis, J. R. y G. G. Koch (1977). «The measurement of observer agreement for categorical data.», *Biometrics*, **33**(1), 159–174.

Leacock, Claudia, Martin Chodorow y George A. Miller (1998). «Using Corpus Statistics and WordNet Relations for Sense Identification», *Computational Linguistics*, **24**(1), 147–165.

Lee, Lillian (1997). «Similarity-based approaches to natural language processing», *CoRR*, **cmp-lg/9708011**.

Lee, Lillian (1999). «Measures of distributional similarity», en *ACL*.

Lesk, Michael (1986). «Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone», en *Proceedings of the 1986 SIGDOC Conference, Association for Computing Machinery*, págs. 24–26, Toronto, Canada.

Lewis, D. y M. Ringuette (1994). «A comparison of two learning algorithms for text categorization», In Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval.

Lin, Dekang (1997). «Using syntactic dependency as local context to resolve word sense ambiguity», en *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistic and the 8th Conference of the European Chapter of the Association for Computational Linguistics*, págs. 64–71.

Lin, Dekang (1998a). «Automatic retrieval and clustering of similar words», en *COLING-ACL*, págs. 768–774.

Lin, Dekang (1998b). «An information-theoretic definition of similarity.», en *ICML*, págs. 296–304.

Lin, Dekang y Patrick Pantel (2001). «Discovery of inference rules for question answering», *Natural Language Engineering*, **7**, 343–360.

Lin, Jianhua (1991). «Divergence measures based on the shannon entropy», *IEEE Transactions on Information Theory*, **37**(1), 145–.

Litkowski, Ken (2004a). «Senseval-3 task: Automatic labeling of semantic roles», en Rada Mihalcea y Phil Edmonds, editores, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, págs. 9–12, Association for Computational Linguistics, Barcelona, Spain.

Litkowski, Ken (2004b). «Senseval-3 task: Word sense disambiguation of wordnet glosses», en Rada Mihalcea y Phil Edmonds, editores, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, págs. 13–16, Association for Computational Linguistics, Barcelona, Spain.

Litkowski, Kenneth C. y Orin Hargraves (2007). «Semeval-2007 task 06: Word-sense disambiguation of prepositions», en *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, págs. 24–29, Association for Computational Linguistics, Prague, Czech Republic.

Llopis, Fernando (2003). «Ir-n: Un sistema de recuperación de información basado en pasajes», Tesis Doctoral. Departamen-

to de Lenguajes y Sistemas Informáticos, Universidad de Alicante.

Luk, A. (1995). «Statistical sense disambiguation with relatively small corpora using dictionary definitions», *Proceedings of the 33rd Meetings of the Association for Computational Linguistics (ACL-95). Cambridge, M.A., 1995*, págs. 181–188.

Magnini, B. y G. Cavaglia (2000). «Integrating subject field codes into WordNet», en *Proceedings of Third International Conference on Language Resources and Evaluation (LREC-2000)*.

Magnini, Bernardo, Danilo Giampiccolo y Alessandro Vallin (2004). «The italian lexical sample task at senseval-3», en Rada Mihalcea y Phil Edmonds, editores, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, págs. 17–20, Association for Computational Linguistics, Barcelona, Spain.

Magnini, Bernardo y C. Strapparava (2000). «Experiments in Word Domain Disambiguation for Parallel Texts», en *Proceedings of the ACL Workshop on Word Senses and Multilinguality*, Hong Kong, China.

Mann, Gideon S. (2006). «Multi-document statistical fact extraction and fusion», Ph.D. Thesis.

Manning, C. D. y H. Schütze (1999). *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, Massachusetts.

Markert, Katja y Malvina Nissim (2007). «Semeval-2007 task 08: Metonymy resolution at semeval-2007», en *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, págs. 36–41, Association for Computational Linguistics, Prague, Czech Republic.

Markov, A. (1971). «Extension of the limit theorems of probability theory to a sum of variables connected in a chain», en *Dynamic Probabilistic Systems, volume 1: Markov Chains.*

Màrquez, Lluis, Mariona Taulé, Antonia Martí, Núria Artigas, Mar García, Francis Real y Dani Ferrés (2004a). «Senseval-3: The spanish lexical sample task», en Rada Mihalcea y Phil Edmonds, editores, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic*

Analysis of Text, págs. 21–24, Association for Computational Linguistics, Barcelona, Spain.

Màrquez, Lluis, Mariona Taulé, Antonia Martí, Mar García, Francis Real y Dani Ferrés (2004b). «Senseval-3: The catalan lexical sample task», en Rada Mihalcea y Phil Edmonds, editores, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, págs. 147–150, Association for Computational Linguistics, Barcelona, Spain.

Màrquez, Lluís, Lluis Villarejo, M. A. Martí y Mariona Taulé (2007). «Semeval-2007 task 09: Multilevel semantic annotation of catalan and spanish», en *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, págs. 42–47, Association for Computational Linguistics, Prague, Czech Republic.

Martínez, David, Eneko Agirre y Lluís Màrquez (2002). «Syntactic features for high precision word sense disambiguation», en *COLING*.

Martínez, D. y E. Agirre (2000). «One sense per collocation and genre/topic variations», en *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Hong Kong.

Martínez, David (2004). «Supervised word sense disambiguation: facing current challenges», Tesis Doctoral. Departamento de Lenguajes y Sisteams Informáticos. Universidad del País Vasco.

McCarthy, Diana, Rob Koeling, Julie Weeds y John A. Carroll (2004). «Finding predominant word senses in untagged text.», en *ACL*, págs. 279–286.

McCarthy, Diana y Roberto Navigli (2007). «Semeval-2007 task 10: English lexical substitution task», en *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, págs. 48–53, Association for Computational Linguistics, Prague, Czech Republic.

McRoy, Susan (1992). «Using multiple knowledge sources for word sense discrimination», *Computational Linguistics*, **18**(1), 1–30.

Universidad de Alicante

Melamed, I. Dan y Philip Resnik (2000). «Tagger evaluation given hierarchical tag sets», *CoRR*, **cs.CL/0008007**.

Mihalcea, Rada, Timothy Chklovski y Adam Kilgarriff (2004a). «The senseval-3 english lexical sample task», en Rada Mihalcea y Phil Edmonds, editores, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, págs. 25–28, Association for Computational Linguistics, Barcelona, Spain.

Mihalcea, Rada y Dan Moldovan (1999). «A Method for word sense disambiguation of unrestricted text», en *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistic*, págs. 152–158, Maryland, Usa.

Mihalcea, Rada y Dan I. Moldovan (2001). «extended wordnet: Progress report», en *in Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, págs. 95–100.

Mihalcea, Rada, Vivi Năstase, Timothy Chklovski, Doina Tătar, Dan Tufiş y Florentina Hristea (2004b). «An evaluation exercise for romanian word sense disambiguation», en Rada Mihalcea y Phil Edmonds, editores, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, págs. 29–32, Association for Computational Linguistics, Barcelona, Spain.

Miller, George y Walter Charles (1991). «Contextual correlates of semantic similarity», en *Language and Cognitive Processes*, págs. 1–28.

Miller, George A. (1995). «Wordnet: A lexical database for english.», *Commun. ACM*, **38**(11), 39–41.

Mitchell, Tom M. (1997a). «Machine learning», en *McGraw Hill*.

Mitchell, Tom M. (1997b). «Machine learning meets natural language», en *EPIA*, pág. 391.

Molina, Antonio, Ferran Pla y Encarna Segarra (2002). «A hidden markov model approach to word sense disambiguation», en *IBERAMIA*, págs. 655–663.

Montoyo, Andrés (2002). «Desambiguación léxica mediante marcas de especificidad», Tesis Doctoral. Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante.

Montoyo, Andrés, Sonia Vázquez y German Rigau (2003). «Método de desambiguación léxica basada en el recurso léxico Dominios Relevantes», *Procesamiento del Lenguaje Natural*, **31**, 141–148.

Mooney, Raymond (1996). «Comparative experiments on disambiguating word sense: An illustration of the role of bias in machine learning», en *Proceedings of the 1st Conference on Empirical Methods in Natural Language Processing*, págs. 82–91.

MUC (1995). «Proceedings of the 6th message understanding conference (muc-6)», Morgan Kaufmann, San Mateo, CA.

MUC (1998). «Proceedings of the 7th message understanding conference (muc-7)», Morgan Kaufmann, San Mateo, CA.

Navarro, Borja, Lorenza Moreno, Sonia Vázquez, Fernando Llopis, Andrés Montoyo y Miguel Angel Varó (2004). «Improving interaction with the user in cross-language question answering through relevant domains and syntactic semantic patterns», en *CLEF*, págs. 334–342.

Navigli, Roberto, Kenneth C. Litkowski y Orin Hargraves (2007). «Semeval-2007 task 07: Coarse-grained english all-words task», en *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, págs. 30–35, Association for Computational Linguistics, Prague, Czech Republic.

Ng, Hwee Tou y Yee Seng Chan (2007). «Semeval-2007 task 11: English lexical sample task via english-chinese parallel text», en *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, págs. 54–58, Association for Computational Linguistics, Prague, Czech Republic.

Nigam, Kamal y Rayid Ghani (2000). «Analyzing the effectiveness and applicability of co-training», en *CIKM*, págs. 86–93.

Niles, Ian y Adam Pease (2001). «Towards a standard upper ontology», en *FOIS*, págs. 2–9.

Niles, Ian y Adam Pease (2003). «Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology», en *IKE*, págs. 412–416.

Orhan, Zeynep, Emine Çelik y Demirgüç Neslihan (2007). «Semeval-2007 task 12: Turkish lexical sample task», en

Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), págs. 59–63, Association for Computational Linguistics, Prague, Czech Republic.

Pantel, Patrick y Dekang Lin (2002). «Document clustering with committees», en *SIGIR*, págs. 199–206.

Patwardhan, Siddharth, Satanjeev Banerjee y Ted Pedersen (2003). «Using measures of semantic relatedness for word sense disambiguation.», en *CICLing*, págs. 241–257.

Pedersen, Ted (1997). «Naive mixes for word sense disambiguation», en *AAAI/IAAI*, pág. 841.

Pedersen, Ted (2002). «Assessing system agreement and instance difficulty in the lexical sample tasks of senseval-2», *CoRR*, **cs.CL/0205068**.

Pedersen, Ted y Rebecca Bruce (1997). «Distinguishing word senses in untagged text», en *Proceedings of the 2th Conference on Empirical Methods in Natural Language Processing*, págs. 197–207.

Pekar, V. y M. Krkoska (2003). «Weighting distributional features for automatic semantic classification of words», en *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03)*, págs. 369–373.

Pekar, Viktor, Ruslan Mitkov, Dimitar Blagoev y Andrea Mulloni (2006). «Finding translations for low-frequency words in comparable corpora», *Machine Translation*, **20**(4), 247–266.

Peñas, Anselmo, Álvaro Rodrigo, Valentín Sama y Felisa Verdejo (2006a). «Overview of the answer validation exercise 2006», en *CLEF*, págs. 257–264.

Peñas, Anselmo, Álvaro Rodrigo y Felisa Verdejo (2006b). «Sparte, a test suite for recognising textual entailment in spanish», en *CICLing*, págs. 275–286.

Pianta, Emanuele, Luisa Bentivogli y Christian Girardi (2002). «Multiwordnet: developing an aligned multilingual database», en *In Proceedings of the First International Conference on Global WordNet*, Mysore (India).

Pradhan, Sameer, Edward Loper, Dmitriy Dligach y Martha Palmer (2007). «Semeval-2007 task-17: English lexi-

cal sample, srl and all words», en *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, págs. 87–92, Association for Computational Linguistics, Prague, Czech Republic.

Preiss, Judita y Anna Korhonen (2004). «Wsd for subcategorization acquisition task description», en Rada Mihalcea y Phil Edmonds, editores, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, págs. 33–36, Association for Computational Linguistics, Barcelona, Spain.

Pustejovsky, James (1991). «The generative lexicon», *Computational Linguistics*, **17**(4), 409–441.

Quillian (1986). «Induction of decision trees.», en *Machine Learning*, págs. 81–106.

Quillian (1993). «C4.5. programs for machine learning», en Morgan Kaufmann, editor, *Machine Learning*.

Quinlan, Ross (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann.

Rada, R., H. Mili, E. Bicknell y M. Blettner (1989). «Development an Application of a Metric on Semantic Nets», *IEEE Transactions on Systems, Man and Cybernetics*, **19**(1), 17–30.

Ramakrishnan, Ganesh, B. Prithviraj y Pushpak Bhattacharya (2004). «A gloss-centered algorithm for disambiguation», en Rada Mihalcea y Phil Edmonds, editores, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, págs. 217–221, Association for Computational Linguistics, Barcelona, Spain.

Rao, C. R. (1982). «Diversity: Its measurement, decomposition, apportionment and analysis», en *Sankya: The Indian Journal of Statistics*, págs. 1–22.

Ratnaparkhi, Adwait (1998). «Maximum entropy models for natural language ambiguity resolution», Tesis Doctoral. Universidad de Pensilvania.

Ravin, Yael y Claudia Leacock (2001). «Polysemy theoretical and computational approaches», en *Oxford University Press*.

Rayson, P. y R. Garside (2000). «Comparing corpora using frequency profiling.», en *In proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*, págs. 1–6.

Resnik, Philip (1993). *Selection and Information: A Class-based Approach to Lexical Relationships*, Tesis Doctoral, University of Pennsylvania.

Resnik, Philip (1995b). «Using information content to evaluate semantic similarity in a taxonomy», en *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, págs. 448–453.

Rigau, German (1998). «Automatic acquisition of lexical knowledge from mrds», Tesis Doctoral. Departamento de Lenguajes y Sistemas Informáticos, Universidad Politécnica de Cataluña.

Rivest, Ronald (1987). «Learning decision lists», *Machine Learning*, **2**, 229–246.

Rosso, Paolo, Francesco Masulli, Davide Buscaldi, Ferran Pla y Antonio Molina (2003). «Automatic noun sense disambiguation», en *CICLing*, págs. 273–276.

Rosso, Paolo, Antonio Molina, Ferran Pla, Daniel Jiménez y Vicente Vidal (2004). «Information retrieval and text categorization with semantic indexing», en *CICLing*, págs. 596–600.

Ruge, Gerda (1992). «Experiments on linguistically-based term associations», *Inf. Process. Manage.*, **28**(3), 317–332.

Rus, Vasile (2004). «A first evaluation of logic form identification systems», en Rada Mihalcea y Phil Edmonds, editores, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, págs. 37–40, Association for Computational Linguistics, Barcelona, Spain.

Russell, Stuart J. y Peter Norvig (1995). «A modern, agent-oriented approach to introductory artificial intelligence», *SIGART Bulletin*, **6**(2), 24–26.

S., Small (1980). «Word expert parsing: A theory of distributed word-based natural language understanding», .

Universidad de Alicante

Sanderson, M. (1994). «Word sense disambiguation and information retrieval», en *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, págs. 142–151.
Schütze, H. (1998). «Automatic word sense discrimination», *Computational Linguistics*, **24**(1), 97–123.
Schütze, H. y J. Pedersen (1995). «Information retrieval based on word senses», en *Proceedings of 4th Annual Symposium on Document Analysis and Information Retrieval*, págs. 161–175.
Small, Steven y Charles Rieger (1982). *Parsing and comprehending with word experts (a theory and its realization)*, págs. 89–147, Lawrence Erlbaum and associates, Hillsdale, NJ, Wendy Lenhert and Martin Ringle ed.
Snyder, Benjamin y Martha Palmer (2004). «The english all-words task», en Rada Mihalcea y Phil Edmonds, editores, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, págs. 41–43, Association for Computational Linguistics, Barcelona, Spain.
Stetina, J., S. Kurohashi y M.Ñagao (1998). «General word sense disambiguation method based on full sentencial context», en *Proceedings of Usage of WordNet in Natural Language Processing. COLING-ACL Workshop*, Montreal, Canada.
Stevenson, Mark y Yorick Wilks (2001). «The interaction of knowledge sources in word sense disambiguation.», *Computational Linguistics*, **27**(3), 321–349.
Strapparava, Carlo, Massimiliano Gliozzo y Claudio Giuliano (2004). «Pattern abstraction and term similarity for word sense disambiguation: Irst at senseval-3», en *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, págs. 229–234, Association for Computational Linguistics.
Strapparava, Carlo y Rada Mihalcea (2007). «Semeval-2007 task 14: Affective text», en *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, págs. 70–74, Association for Computational Linguistics, Prague, Czech Republic.

Suárez, Armando y Manuel Palomar (2002). «A maximum entropy-based word sense disambiguation system», en *COLING*.

Suárez, Armando (2004). «Resolución de la ambigüedad semántica de las palabras mediante modelos de probabilidad de máxima entropía», Tesis Doctoral. Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante.

Towell, Geoffrey G. y Ellen M. Voorhees (1998). «Disambiguating highly ambiguous words», *Computational Linguistics*, **24**(1), 125–145.

Turney, Peter D. (2001). «Mining the web for synonyms: Pmi-ir versus lsa on toefl», en *ECML*, págs. 491–502.

Turney, Peter D. (2004). «Human-level performance on word analogy questions by latent relational analysis», *CoRR*, **abs/cs/0412024**.

Ulivieri, Marisa, Elisabetta Guazzini, Francesca Bertagna y Nicoletta Calzolari (2004). «Senseval-3: The italian all-words task», en Rada Mihalcea y Phil Edmonds, editores, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Association for Computational Linguistics, Barcelona, Spain.

Valdivia, M. T. Martín, Alfonso Ureña López y Manuel García Vega (2002). «Resolución de la ambigüedad mediante redes neuronales», *Procesamiento del Lenguaje Natural*, **29**, 39–45.

Vapnik, Vladimir (1998). «Statistical learning theory», en *New York USA: John Wiley*.

Vasilescu, Florentina, Philippe Langlais y Guy Lapalme (2004). «Evaluating variants of the Lesk approach for disambiguating words», *Proceedings of the Conference on language Resources and Evaluation (LREC)*, págs. 633–636.

Vázquez, Sonia, Zornitsa Kozareva y Andrés Montoyo (2006). «Textual entailment beyond semantic similarity information», en *MICAI*, págs. 900–910.

Vázquez, Sonia, Andrés Montoyo y Zornitsa Kozareva (2007). «Word sense disambiguation using extended relevant domains resource», en *IC-AI*, págs. 823–828.

Universidad de Alicante

Vega, Manuel García, María Teresa Martín-Valdivia y Luis Alfonso Ureña (2003). «Aprendizaje competitivo LVQ para la desambiguación léxica», *Procesamiento del Lenguaje Natural*, **31**, 125–132.

Verhagen, Marc, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz y James Pustejovsky (2007). «Semeval-2007 task 15: Tempeval temporal relation identification», en *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, págs. 75–80, Association for Computational Linguistics, Prague, Czech Republic.

Veronis, Jean y Nancy Ide (1990). «Word Sense Disambiguation with very large neural networks extracted from machine readable dictionaries», en *Proceedings of the 13th International Conference on Computational Linguistics, COLING´90,volume 2*, págs. 389–394, Helsinki, Finland.

Vossen, P. (1998). «The Restructured Core WordNets in EutoWordNet: Subset1.», en *Deliverable D014, D015, WP3, WP4*, EuroWordNet LE2-4003.

Vossen, Piek, German Rigau, Iñaki Alegría, Eneko Agirre, David Farwell y Manuel Fuentes (2006). «Meaningful results for information retrieval in the MEANING project», *Proceedings of the 3rd Global WordNet Conference*.

Vázquez, Sonia, Zornitsa Kozareva y Andrés Montoyo (2007). «How context and semantic information can help a machine learning system», en *Mexican International Conference on Artificial Intelligence*, Lecture Notes in Computer Science, págs. 996–1003, Aguascalientes (Mexico).

Widdows, Dominic y Stanley Peters (2003). «Word vectors and quantum logic experiments with negation and disjunction», en *Proceedings of Mathematics of Language*, págs. 141–154.

Wilks, Yorick (1972). «Grammar, meaning and the machine analysis of language.», en *London:Routledge*.

Wilks, Yorick y Mark Stevenson (1998). «Word sense disambiguation using optimised combinations of knowledge sources», en *COLING-ACL*, págs. 1398–1402.

WITTEN, IAN H. y EIBE FRANK (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann.

YAROWSKY, DAVID (1992). «Word sense disambiguation using statistical models of Roget´s categories trained on large corpora», en *Proceedings of the 14th International Conference on Computational Linguistic, COLING´92*, págs. 454–460, Nantes, France.

YAROWSKY, DAVID (1993). «One sense per collocation», en *Proceedings of the DARPA Workshop on Human Language Technology*, págs. 266–271, Princenton, NJ.

YAROWSKY, DAVID (1994a). «A comparison of corpus-based techniques for restoring accents in Spanish and French text», en *Proceedings of the 2th Annual Workshop on Very Large Text Corpora*, págs. 19–32.

YAROWSKY, DAVID (1994b). «Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French», en *Proceedings of the 32th Annual Meeting of the Association for Computational Linguistic*, págs. 88–95.

YAROWSKY, DAVID (1995). «Unsupervised word sense disambiguation rivaling supervised methods», en *Proceedings of the 33th Annual Meeting of the Association for Computational Linguistic*, págs. 189–196.

YAROWSKY, DAVID (1996). «Homograph disambiguation in text-to-speech synthesis», .

YAROWSKY, DAVID (2000a). «Hierarchical decision lists for word sense disambiguation», *Computers and the Humanities*, **34**, 1–2.

YAROWSKY, DAVID (2000b). «Hierarchical decision lists for word sense disambiguation», en *Computers and the Humanities*, págs. 179–186.

YU, CLARA, JOHN CUADRADO, MACIEJ CEGLOWSKI y J. SCOTT PAYNE (2004). «Patterns in unstructured data, discovery, aggregation and visualization», en *Presentation to the Andrew W. Mellon Foundation*.

Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Reunido el Tribunal que suscribe en el día de la fecha acordó otorgar, por _____ a la Tesis Doctoral de Don/Doña. Sonia Vázquez Pérez la calificación de _____.

Alicante _____ de _____ de _____

El Secretario,

El Presidente,

UNIVERSIDAD DE ALICANTE
Comisión de Doctorado

La presente Tesis de D. Sonia Vázquez Pérez ha sido registrada con el nº _____ del registro de entrada correspondiente.

Alicante _____ de _____ de _____

El Encargado del Registro,

Universitat d'Alacant
Universidad de Alicante

La defensa de la tesis doctoral realizada por D/D^{a} Sonia Vázquez Pérez se ha realizado en las siguientes lenguas: _____ y _____ , lo que unido al cumplimiento del resto de requisitos establecidos en la Normativa propia de la UA le otorga la mención de "Doctor Europeo".

Alicante, _____ de _____ de _____

EL SECRETARIO EL PRESIDENTE

Universitat d'Alacant
Universidad de Alicante